# THE ANNALS *of* MATHEMATICAL STATISTICS

THE ANNALS OF MATHEMATICAL STATISTICS IS AFFILIATED WITH THE AMERICAN STATISTICAL ASSOCIATION AND IS DEVOTED TO THE THEORY AND APPLICATION OF MATHEMATICAL STATISTICS



Volume VII, 1936

PUBLISHED QUARTERLY
ANN ARBOR, MICHIGAN

*Four Dollars per annum*

*Made in United States of America*

Address: ANNALS OF MATHEMATICAL STATISTICS
Post Office Box 171, Ann Arbor, Michigan

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD.

# THE ANNALS *of* MATHEMATICAL STATISTICS

THE ANNALS OF MATHEMATICAL STATISTICS IS AFFILIATED WITH THE AMERICAN STATISTICAL ASSOCIATION AND IS DEVOTED TO THE THEORY AND APPLICATION OF MATHEMATICAL STATISTICS



Volume VII, Number 1
MARCH, 1936

PUBLISHED QUARTERLY
ANN ARBOR, MICHIGAN

*Four Dollars per annum*

*Made in United States of America*

Address: ANNALS OF MATHEMATICAL STATISTICS
Post Office Box 171, Ann Arbor, Michigan

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD.

# ON THE FREQUENCY FUNCTION OF $xy$

By Cecil C. Craig

Given the distribution function of $x$ and $y$, what can be said of the distribution of the product $xy$? The author has had two inquiries during the last two years, one from an investigator in business statistics and the other from a psychologist, concerning the probable error of the product of two quantities, each of known probable error. There seems to be very little in the literature of mathematical statistics on this question.

If $x$ and $y$ are independent and are each distributed according to the same normal frequency law, it is well known that the distribution function of

$$\bar{z} = \frac{x - m_x}{\sigma_x} \cdot \frac{y - m_y}{\sigma_y}$$

is

$$\frac{1}{\pi} K_0(\bar{z}) \text{ ,}^1$$

in which $K_0(\bar{z})$ is the Bessel function of the second kind of a purely imaginary argument of zero order.[2] If $x$ and $y$ are independent and are each distributed according to a logarithmic normal frequency law, it has been pointed out that the product, $(x - a)$ $(y - b)$, in which $a$ and $b$ are the upper (or lower) limits of the range for $x$ and $y$ respectively, is distributed according to a law of the same type.[3] In both cases the special choice of origins greatly simplifies the problem.

In the present discussion it will be assumed that $x$ and $y$ are distributed normally. It will appear that the distribution of $xy$ is a function of $r_{xy}$, the coefficient of correlation between $x$ and $y$, and of the parameters,

$$\rho_1 = \frac{m_1}{\sigma_1} = \frac{m_x}{\sigma_x} \quad \text{and} \quad \rho_2 = \frac{m_2}{\sigma_2} = \frac{m_y}{\sigma_y} ,$$

which are proportional to the reciprocals of the coefficients of variation. The chief difficulty arises when $\rho_1$ and $\rho_2$ are small so that zero values of $xy$ occur

[1] J. Wishart and M. S. Bartlett: The Distribution of Second Order Moment Statistics in a Normal System; Proceedings of the Cambridge Philosophical Society, Vol. XXVIII (1932), pp. 455–459.

[2] G. N. Watson: A Treatise on the Theory of Bessel Functions; Cambridge University Press (1922), p. 78.

[3] P. T. Yuan: On the Logarithmic Frequency Distribution and the Semi-logarithmic Frequency Surface; Annals of Mathematical Statistics, Vol. 4 (1933), pp. 46, 47.

for values of $x$ and $y$ well within their respective ranges of variation. (If $\rho_1$ and $\rho_2$ are large, practically one may exclude zero values of $x$ and $y$ from consideration. The author hopes to present an investigation of this case soon.) It is the object of the present paper to study the rather unusual frequency function that arises in this situation. It will first be assumed that $x$ and $y$ are independent ($r_{xy} = r = 0$). Then it will be shown that the distribution function when $r \neq 0$ is readily derived from that arrived at in the special case.

We can find the moment generating function of $xy$ without difficulty. We have,

$$M_{xy}(\vartheta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}-\frac{(y-m_2)^2}{2\sigma_2^2}} e^{xy\vartheta}\, dx\, dy$$

$$= \frac{e^{[(\sigma_1^2 m_2^2+\sigma_2^2 m_1^2)\vartheta^2+2m_1m_2\vartheta]/2(1-\sigma_1^2\sigma_2^2\vartheta^2)}}{(1-\sigma_1^2\,\sigma_2^2\,\vartheta^2)^{1/2}}.$$

Setting, for convenience,

$$z = \frac{xy}{\sigma_1\sigma_2},$$

this can be written,

$$M_z(\vartheta) = \frac{e^{[(\rho_1^2+\rho_2^2)\vartheta^2+2\rho_1\rho_2\vartheta]/2(1-\vartheta^2)}}{(1-\vartheta^2)^{1/2}}. \tag{1}$$

This choice of variable and of parameters will be adhered to in the sequel.

On expanding $\log M_z(\vartheta)$ in powers of $\vartheta$, we get for the semi-invariants (of Thiele),

$$\begin{aligned}\lambda_{2k+1:z} &= (2k+1)!\,\rho_1\rho_2, \qquad k = 0, 1, 2, \cdots \\ \lambda_{2k:z} &= \frac{(2k)!}{2}(\rho_1^2+\rho_2^2) + (2k-1)!, \qquad k = 1, 2, \cdots.\end{aligned} \tag{2}$$

These give for the mean and variance of $xy$,

$$M_{xy} = m_1m_2$$

$$\sigma_{xy}^2 = \sigma_1^2 m_2^2 + \sigma_2^2 m_1^2 + \sigma_1^2\sigma_2^2.$$

For the standard semi-invariants of $z$ (or of $xy$), we have,

$$\xi_{2k+1:z} = \frac{\lambda_{2k+1:z}}{\lambda_{2:z}^{\frac{2k+1}{2}}} = \frac{(2k+1)!\,\rho_1\rho_2}{(\rho_1^2+\rho_2^2+1)^{\frac{2k+1}{2}}},$$

$$\xi_{2k:z} = \frac{\lambda_{2k:z}}{\lambda_{2:z}^{k}} = \frac{(2k-1)!\,[k(\rho_1^2+\rho_2^2)+1]}{(\rho_1^2+\rho_2^2+1)^k}.$$

Taking,

$$\xi_3 = \frac{6\,\rho_1\rho_2}{(\rho_1^2 + \rho_2^2 + 1)^{3/2}},$$

as a measure of skewness, it is easy to verify that

$$|\,\xi_3\,| \leqq \frac{2}{3}\sqrt{3}\,.$$

For either $\rho_1 = 0$ or $\rho_2 = 0$, the distribution is symmetrical about its mean which then falls at the origin.

For the excess or kurtosis, we have,

$$\xi_4 = \frac{6\,[2(\rho_1^2 + \rho_2^2) + 1]}{(\rho_1^2 + \rho_2^2 + 1)^2} \leqq 6\,.$$

Thus the skewness is never great and becomes small with increasing $\rho_1$ or $\rho_2$. The excess also becomes small with increasing $\rho_1$ or $\rho_2$, but it can be very large for small values of these parameters, attaining its maximum of 6 for $\rho_1 = \rho_2 = 0$. But, as it will appear below, the distribution function always becomes infinite in a logarithmic manner at the origin. (We have already seen, as must obviously be the case, that moments of all orders exist.) It is to be noted, too, that for any given $\rho_1$ and $\rho_2$, $\xi_{2k}$ increases without limit with increasing $k$, and that the same is true of $\xi_{2k+1}$ if neither $\rho_1$ nor $\rho_2$ is zero.

Turning now to the derivation of the actual frequency function of $z$, we set $w = xy$; then for any given $x$, $y = w/x$, $dy = \frac{dw}{x}$ if $x > 0$, and $dy = -\frac{dw}{x}$ if $x < 0$. These values are substituted into $\varphi_1(x)\,\varphi_2(y)dx\,dy$, in which $\varphi_1(x)$ and $\varphi_2(y)$ are the frequency functions of $x$ and $y$ respectively, and the resulting expression is integrated over all values of $x$, giving for the frequency function of $w$:

$$F(w) = \frac{e^{-\left(\frac{m_1^2}{2\,\sigma_1^2} + \frac{m_2^2}{2\,\sigma_2^2}\right)}}{2\pi\sigma_1\sigma_2}\left[\int_0^\infty \Phi(w,\,x)\,\frac{dx}{x} - \int_{-\infty}^0 \Phi(w,\,x)\,\frac{dx}{x}\right]$$

in which,

$$\Phi(w,\,x) = e^{-(\sigma_2^2x^4 - 2m_1\sigma_2^2x^3 - 2m_2\sigma_1^2wx + \sigma_1^2w^2)/2\sigma_1^2\sigma_2^2x^2}\,.$$

Again setting $z = \frac{xy}{\sigma_1\sigma_2}$, and introducing the parameters $\rho_1$ and $\rho_2$, this reduces to,

$$F(z) = \frac{e^{-\frac{(\rho_1^2+\rho_2^2)}{2}}}{2\pi}[\psi_1(z) - \psi_2(z)]\,, \tag{4}$$

in which,

$$\psi_1(z) = \int_0^\infty e^{-\left(\frac{x^2}{2}-\rho_1 x-\rho_2\frac{z}{x}+\frac{z^2}{2x^2}\right)}\frac{dx}{x}, \tag{5}$$

and $\psi_2(z)$ is the integral of the same function over the interval $(-\infty, 0)$.

Now writing,

$$\psi_1(z) = \int_0^\infty e^{-\frac{x^2}{2}-\frac{z^2}{2x^2}}\frac{e^{\rho_1 x+\rho_2\frac{z}{x}}}{x}\,dx, \tag{6}$$

we note that

$$\frac{e^{\rho_1 x+\rho_2\frac{z}{x}}}{x}$$

can be expanded in a Laurent series in powers of $x$ for all values of $x$ except zero. In this expansion the coefficient of $x^{r-1}$, $r \geqq 1$, is $\frac{\rho_1^r}{r!}\sum_r(\rho_1\rho_2 z)$, in which

$$\begin{aligned}\sum\nolimits_r(\rho_1\rho_2 z) &= 1+\frac{\rho_1\rho_2 z}{r+1}+\frac{(\rho_1\rho_2 z)^2}{(r+2)^{(2)}\,2!}+\frac{(\rho_1\rho_2 z)^3}{(r+3)^{(3)}\,3!}+\cdots,\\ ((r+k)^{(k)} &= (r+k)(r+k-1)\cdots(r+1)).\end{aligned} \tag{7}$$

We may note parenthetically that

$$\frac{\rho_1^r}{r!}\sum\nolimits_r(\rho_1\rho_2 z) = \left(\frac{\rho_1}{\rho_2 z}\right)^{\frac{r}{2}} I_r(2\sqrt{\rho_1\rho_2 z}),$$

in which $I_r(x)$ is the Bessel function of the first kind with a purely imaginary argument.[4]

The coefficient of $x^{-r-1}$, $r \geqq 0$, is $\frac{z^r\rho_2^r}{r!}\sum_r(\rho_1\rho_2 z)$.

Setting now,

$$\sum_{n=-\infty}^{\infty} f_n(x) = \frac{e^{\rho_1 x+\rho_2\frac{z}{x}}}{x},$$

we substitute this series in (6) and seek to justify the expansion it gives for $\psi_1(z)$ obtained by term by term integration. We write,

$$\psi_1(z) = \int_0^1 e^{-\frac{x^2}{2}-\frac{z^2}{2x^2}}\sum f_n(x)\,dx + \int_1^\infty e^{-\frac{x^2}{2}-\frac{z^2}{2x^2}}\sum f_n(x)\,dx.$$

[4] Watson, loc. cit., p. 77.

For $z > 0$, $\rho_1\rho_2 > 0$, the terms of $\sum f_n(x)$ are all $> 0$. Then the convergence of

$$\sum \int_0^1 e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx$$

is sufficient to allow term by term integration in the first integral. In the second integral we observe that $\sum f_n(x)$ converges uniformly in every fixed interval $1 \leqq x \leqq a$. Then term by term integration is permissible here if

$$\sum \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx$$

is convergent.[5] It is evident, then, that it will be sufficient to establish the convergence of

$$\sum \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2x^2}} f_n(x)\, dx\,.$$

If either or both $z < 0$ or $\rho_1\rho_2 < 0$, it will be easily seen that the series involved are still absolutely convergent which is sufficient.

Now using the definition of the Bessel function of a purely imaginary argument of the second kind,

$$K_\nu(z) = \frac{1}{2}\left(\frac{z}{2}\right)^\nu \int_0^\infty e^{-\tau - \frac{z^2}{4\tau}} \frac{d\tau}{\tau^{\nu+1}}\,,$$ [6]

it is easy to derive the relation,

$$K_{\frac{n-1}{2}}(z) = z^{\frac{n-1}{2}} \int_0^\infty e^{-\frac{x^2}{2} - \frac{z^2}{2k^2}} \frac{dx}{x^n}\,.$$

Remembering that $K_\nu(z) = K_{-\nu}(z)$, we have for our expansion,

$$\psi_1(z) = \sum\nolimits_0 K_0 + (\rho_1 + \rho_2)\, z^{1/2} \sum\nolimits_1 K_{\frac{1}{2}} + (\rho_1^2 + \rho_2^2)\, \frac{z}{2!} \sum\nolimits_2 K_1$$

$$+ (\rho_1^3 + \rho_2^3)\, \frac{z^{\frac{3}{2}}}{3!} \sum\nolimits_3 K_{\frac{3}{2}} + \cdots$$

in which the argument for all the $\sum$-functions is $\rho_1\rho_2\, z$, and for all the $K$-functions is $z$.

---

[5] T. J. I'a Bromwich: An Introduction to the Theory of Infinite Series; Macmillan & Co., London, 2nd edition (1926), pp. 496 and 500.

[6] Watson, loc. cit., pp. 78 and 183.

But we may as well add to this the expansion of $-\psi_2(z)$, which may be written,

$$\int_0^\infty e^{-\frac{x^2}{2}-\frac{z^2}{2x^2}} \frac{e^{-\rho_1 x-\rho_2\frac{z}{x}}}{x}\,dx\,,$$

and obtain the expansion,

$$F(z) = \frac{e^{-\frac{\rho_1^2+\rho_2^2}{2}}}{\pi}\left[\sum\nolimits_0 K_0 + (\rho_1^2+\rho_2^2)\frac{z}{2!}\sum\nolimits_2 K_1 + (\rho_1^4+\rho_2^4)\frac{z^2}{4!}\sum\nolimits_4 K_2 + \cdots\right],$$

the convergence of which we will examine. But it must be noted that the terms arising from the expansion of

$$\frac{e^{\rho_1 x+\rho_2\frac{z}{x}}}{x} \quad\text{and}\quad \frac{e^{-\rho_1 x-\rho_2\frac{z}{x}}}{x}$$

which contribute to the expansion of $F(z)$ as just written are those of the forms,

$$\frac{\rho_1^{2i}}{(2i)!}\sum\nolimits_{2i} \quad\text{and}\quad \frac{\rho_2^{2i}\,z^{2i}}{(2i)!}\sum\nolimits_{2i}\,.$$

Hence the expansion as written is valid in any case only for $z > 0$. For $z \gtrless 0$, we may write however,

$$(8)\qquad F(z) = \frac{e^{-\frac{\rho_1^2+\rho_2^2}{2}}}{\pi}\left[\sum\nolimits_0 K_0 + (\rho_1^2+\rho_2^2)\frac{|z|}{2!}\sum\nolimits_2 K_1 + (\rho_1^4+\rho_2^4)\frac{z^2}{4!}\sum\nolimits_4 K_2 \right.$$
$$\left. + (\rho_1^6+\rho_2^6)\frac{|z|^3}{6!}\sum\nolimits_6 K_3 + \cdots\right],$$

in which the arguments for the $\sum$ and $K$-functions are the same as before.

Let us consider now the question of the convergence of (8), first in the case that $z > 0$. We set

$$c_\nu = \frac{z^\nu K_\nu}{\nu!}\bigg/\frac{z^{\nu-1}K_{\nu-1}}{(\nu-1)!}\,.$$

Then from the relation,

$$(9)\qquad K_{\nu-1} - K_{\nu+1} = -\frac{2\nu}{z}K_\nu\,,\text{[7]}$$

we readily derive,

$$\frac{z^2}{(\nu+1)^{(2)}} = c_\nu\left(c_{\nu+1} - \frac{2\nu}{\nu+1}\right).$$

[7] Watson, loc. cit., p. 79.

For $z > 0$, the left hand member and $c_\nu$ are both $> 0$. Thus

$$c_{\nu+1} - \frac{2\nu}{\nu + 1} > 0 .$$

Then let

$$c_{\nu+1} = \frac{2\nu}{\nu + 1} + \delta_{\nu+1} , \qquad \delta_{\nu+1} > 0 ,$$

and we have,

$$\frac{z^2}{(\nu + 1)^{(2)}} > \left(2 - \frac{2}{\nu}\right) \delta_{\nu+1} = 2\, \delta_{\nu+1} - \frac{2\, \delta_{\nu+1}}{\nu} .$$

It is evident from this that for a given $z > 0$, a $\nu_0$ exists such that $c_\nu < 3$ for $\nu \geqq \nu_0$.
Further since

$$\textstyle\sum_r \leqq e^{|\rho_1 \rho_2 z|}$$

the convergence sought follows for $z > 0$. Since $K$ is an even function of $z$, it is easy to see that (8) is also convergent for $z < 0$. For $z = 0$, the first term possesses a logarithmic discontinuity at the origin.

To calculate ordinates of $F(z)$ there are fairly extensive tables available in Watson's treatise already referred to. These tables may be readily extended by means of the asymptotic formula for $K(z)$ for larger values of $z$, and by means of (9) for larger values of $\nu$. One can rapidly build up tables of $\sum_r(x)$ by means of the easily obtained recursion formula,

$$\textstyle\sum_r(x) = \sum_{r+1}(x) + \dfrac{x}{(r + 2)^{(2)}} \sum_{r+2}(x) .$$

It is unfortunately true that the expansion found for $F(z)$ is very slowly convergent for large values of $\rho_1$ and $\rho_2$.

At the end of this paper are shown three charts of $F(z)$ with the tables of ordinates from which they were made by way of illustrating what such curves look like. (On the second for comparison the broken line is the normal curve of error.)

For $\rho_1 = \rho_2 = r = 0$, we have simply the known result,

$$F(z) = \frac{1}{\pi} K_0 (z) .$$

For $\rho_1 = 1$, $\rho_2 = r = 0$, the curve is symmetrical about its mean (and the origin). Here every $\sum$-function is unity.

For the case in which $\rho_1 = \rho_2 = \frac{1}{2}$, $r = 0$, I first constructed tables of $\sum_i(x)$ for $x = \pm 0.025, \pm 0.05, \pm 0.1$, and by intervals of 0.1 to $\pm 3.0$ for $i = 0, 1, \cdots, 20$. Values of $\sum_0(x)$ and $\sum_2(x)$ for $x = 3.2$ and 3.4 were also used. Not more than five terms of (8) were required to obtain values of $F(z)$ accurate

to five places of decimals. This distribution curve is skew with $M_z = 0.25$ and $\xi_{3:z} = \dfrac{\sqrt{6}}{3}$.

The curves are plotted in standard units with unit total area ($\sigma_z = \sqrt{\rho_1^2 + \rho_2^2 + 1}$). The tables of ordinates are given both in units of $z = \dfrac{xy}{\sigma_1\sigma_2}$ and of $t = \dfrac{z - m_z}{\sigma_z}$.

Turning now to the case in which $r \neq 0$, after some computation, we have for the moment generating function,

$$(10) \qquad M_z(\vartheta) = \frac{e^{\frac{(\rho_1^2+\rho_2^2-2r\rho_1\rho_2)\vartheta^2+2\rho_1\rho_2\vartheta}{2[1-(1+r)\vartheta][1+(1-r)\vartheta]}}}{\sqrt{[1-(1+r)\vartheta][1+(1-r)\vartheta]}}\,.$$

As a check on this result, if we set $r = 1$ and $\rho_1 = \rho_2 = \rho$ in it we get,

$$M_{\frac{x^2}{\sigma^2}}(\vartheta) = \frac{e^{\frac{\rho^2\vartheta}{1-2\vartheta}}}{\sqrt{1-2\vartheta}}\,,$$

which may be readily verified to be the moment generating function of $\dfrac{x^2}{\sigma^2}$ if $x$ is distributed normally with mean $m$ and variance $\sigma^2$ $\left(\rho = \dfrac{m}{\sigma}\right)$.

To obtain the semi-invariants of $z$ in this case, on expanding $\log M_z(\vartheta)$ in powers of $\vartheta$, setting

$$a = \rho_1^2 + \rho_2^2 - 2\rho_1\rho_2 r\,, \qquad b = 2\rho_1\rho_2\,, \qquad c = 1 + r\,, \qquad \text{and} \qquad d = 1 - r\,,$$

we have,

$$(11) \qquad \begin{aligned} \log M_z(\vartheta) &= \frac{a\vartheta^2 + b\vartheta}{2}(1 - c\vartheta)^{-1}(1 + d\vartheta)^{-1} \\ &\quad - \frac{1}{2}[\log(1 - c\vartheta) + \log(1 + d\vartheta)] \\ &= \frac{a\vartheta^2 + b\vartheta}{4}[2 + (c^2 - d^2)\vartheta + (c^3 + d^3)\vartheta^2 + (c^4 - d^4)\vartheta^3 + \cdots] \\ &\quad + \frac{1}{2}\left[(c - d)\vartheta + (c^2 + d^2)\frac{\vartheta^2}{2} + (c^3 - d^3)\frac{\vartheta^3}{3} + \cdots\right], \end{aligned}$$

from which we derive,

$$(12) \qquad \begin{aligned} \lambda_{n:z} &= \frac{n!}{4}[\{c^{n-1} - (-d)^{n-1}\}a + \{c^n - (-d)^n\}b] \\ &\quad + \frac{(n-1)!}{2}\{c^n + (-d)^n\}\,. \end{aligned}$$

In particular,

$$\lambda_{1:z} = \frac{b}{2} + \frac{c-d}{2} = \rho_1\rho_2 + r$$

$$\lambda_{2:z} = a + \frac{c^2 - d^2}{2} \cdot b + \frac{c^2 + d^2}{2} = \rho_1^2 + \rho_2^2 + 2\rho_1\rho_2 r + (1 + r^2)$$

$$\begin{aligned} \lambda_{3:z} &= \frac{3}{2}\left[(c^2 - d^2)\, a + (c^3 + d^3)\, b\right] + c^3 - d^3 \\ &= 6\left[(\rho_1^2 + \rho_2^2)\, r + \rho_1\rho_2\,(1 + r^2)\right] + 2\,r\,(3 + r^2) \\ \lambda_{4:z} &= 6\left[(c^3 + d^3)\, a + (c^4 - d^4)\, b\right] + 3\,(c^4 + d^4) \\ &= 12\,(\rho_1^2 + \rho_2^2)\,(1 + 3\,r^2) + 24\,\rho_1\rho_2 r\,(3 + r^2) + 6\,(1 + 6r + r^4)\,. \end{aligned} \tag{13}$$

Noting that

$$\frac{\partial a}{\partial r} = -\,b\,, \qquad \frac{\partial b}{\partial r} = 0\,, \qquad \frac{\partial c}{\partial r} = 1\,, \qquad \frac{\partial d}{\partial r} = -\,1\,,$$

one can easily demonstrate what seems to be a rather striking property of these semi-invariants, viz.,

$$\frac{\partial \lambda_{n:z}}{\partial r} = n(n-1)\,\lambda_{n-1:z}\,. \tag{14}$$

To gain a notion of the magnitude of the skewness and excess in this case, we form,

$$\xi_{3:z} = \frac{\lambda_{3:z}}{\lambda_{2:z}^{\frac{3}{2}}} \quad \text{and} \quad \xi_{4:z} = \frac{\lambda_{4:z}}{\lambda_{2:z}^{2}}\,.$$

In view of the above property,

$$\frac{\partial \xi_3}{\partial r} = \frac{6\,\lambda_2^2 - 3\,\lambda_3\lambda_1}{\lambda_2^{\frac{5}{2}}}\,.$$

The denominator of this fraction is always $> 0$. The numerator, after some reduction, can be written,

$$\begin{aligned} 6\,[\rho_1^4 + \rho_2^4 - \rho_1^2\rho_2^2\,(1 - r^2) + (\rho_1^2 + \rho_2^2)\,(2 - r^2) \\ + (\rho_1^2 + \rho_2^2 - 2)\,\rho_1\rho_2 r + 1 - r^2]\,. \end{aligned} \tag{15}$$

The first two terms taken together, the third, and the last are all obviously $> 0$. The term,

$$(\rho_1^2 + \rho_2^2 - 2)\,\rho_1\rho_2 r$$

has its maximum value for $|r| = 1$. But for $r = 1$, (15) becomes,

$$\rho_1^4 + \rho_2^4 + \rho_1\rho_2(\rho_1^2 + \rho_2^2) + (\rho_1 - \rho_2)^2 ,$$

and for $r = -1$, it is,

$$\rho_1^4 + \rho_2^4 - \rho_1\rho_2(\rho_1^2 + \rho_2^2) + (\rho_1 + \rho_2)^2 ,$$

both of which expressions are easily seen to be $> 0$.

Thus (15) is always positive and the maximum value of $\xi_{3:z}$ is attained for $r = 1$, the minimum value for $r = -1$. These values are respectively,

$$\frac{6\,(\rho_1 + \rho_2)^2 + 8}{[(\rho_1 + \rho_2)^2 + 2]^{\frac{3}{2}}} \quad \text{and} \quad \frac{-\,6\,(\rho_1 - \rho_2)^2 - 8}{[(\rho_1 - \rho_2)^2 + 2]^{\frac{3}{2}}} ,$$

the absolute value of either being $\leqq 2\sqrt{2}$, which is attained in the first case for $\rho_1 = -\rho_2$ and in the second for $\rho_1 = \rho_2$. It is seen that for high correlation between $x$ and $y$ the skewness of $xy$ can be quite large.

For the excess, we see that

$$\xi_{4:z} = \frac{\lambda_{4:z}}{\lambda_{2:z}^2}$$

attains a value of 12 when $\rho_1 = -\rho_2$, $r = 1$ or when $\rho_1 = \rho_2$, $r = -1$. Since this is such an extraordinary value it does not seem worth while to carry out the extended computation that seems to be required to verify one's surmise that this is the maximum of the absolute value.

Now, to derive the frequency function we proceed as before. We set $z = \dfrac{xy}{\sigma_1\sigma_2}$ and then

$$F(z) = I_1(z) - I_2(z) ,$$

in which,

$$I_1(z) = \frac{1}{2\pi\sqrt{1 - r^2}} \int_0^\infty e^{-\frac{1}{2(1-r^2)}\left[(x-\rho_1)^2 - 2r(x-\rho_1)\left(\frac{z}{x}-\rho_2\right) + \left(\frac{z}{x}-\rho_2\right)^2\right]} \frac{dx}{x} ,$$

and $I_2(z)$ is the integral of the same function over the interval $(-\infty, 0)$.

We can write $I_1(z)$:

$$\frac{e^{-\frac{\rho_1^2 - 2r\rho_1\rho_2 + \rho_2^2}{2(1-r^2)} + \frac{rz}{1-r^2}}}{2\pi\sqrt{1 - r^2}} \int_0^\infty e^{-\frac{1}{2(1-r^2)}\left(x^2 + \frac{z^2}{x^2}\right) + \frac{1}{1-r^2}\left[(\rho_1 - r\rho_2)x + (\rho_2 - r\rho_1)\frac{z}{x}\right]} \frac{dx}{x} .$$

Setting,

$$\frac{x}{\sqrt{1 - r^2}} = u \quad \text{and} \quad \frac{z}{1 - r^2} = \zeta ,$$

this becomes,

$$\text{(16)} \quad \frac{\sqrt{1-r^2}\, e^{-\frac{\rho_1^2 - 2r\rho_1\rho_2 + \rho_2^2}{2(1-r^2)} + \frac{r\zeta}{(1-r^2)^2}}}{2\pi} \times \int_0^\infty e^{-\frac{1}{2}\left(u^2 + \frac{\zeta^2}{u^2}\right)}\, e^{\frac{\rho_1 - r\rho_2}{\sqrt{1-r^2}} u + \frac{\rho_2 - r\rho_1}{\sqrt{1-r^2}} \frac{\zeta}{u}} \frac{du}{u}.$$

But on writing,

$$\frac{\rho_1 - r\rho_2}{\sqrt{1-r^2}} = R_1 \quad \text{and} \quad \frac{\rho_2 - r\rho_1}{\sqrt{1-r^2}} = R_2,$$

the integral in the last expression is of the same form as the $\psi_1(z)$ in the uncorrelated case. It is evident, then, that the distribution function of $\zeta$ can be written,

$$\text{(17)} \quad \begin{aligned} &\frac{\sqrt{1-r^2}}{\pi} e^{-\frac{\rho_1^2 - 2r\rho_1\rho_2 + \rho_2^2}{2(1-r^2)}} e^{\frac{r\zeta}{(1-r^2)^2}} \Big[ \textstyle\sum_0 (R_1R_2\zeta) K_0(\zeta) \\ &+ (R_1^2 + R_2^2) \frac{|\zeta|}{2!} \textstyle\sum_2 (R_1R_2\zeta) K_1(\zeta) + (R_1^4 + R_2^4) \frac{\zeta^2}{4!} \textstyle\sum_4 (R_1R_2\zeta) K_2(\zeta) \\ &+ (R_1^6 + R_2^6) \frac{|\zeta|^3}{6!} \textstyle\sum_6 (R_1R_2\zeta) K_3(\zeta) + \cdots \Big], \end{aligned}$$

and is essentially of the form of $F(z)$, reached when $r = 0$, multiplied by an exponential function.

Frequency curves for $xy$ (in standard units) are given in Fig. 1, Fig. 2 and Fig. 3.

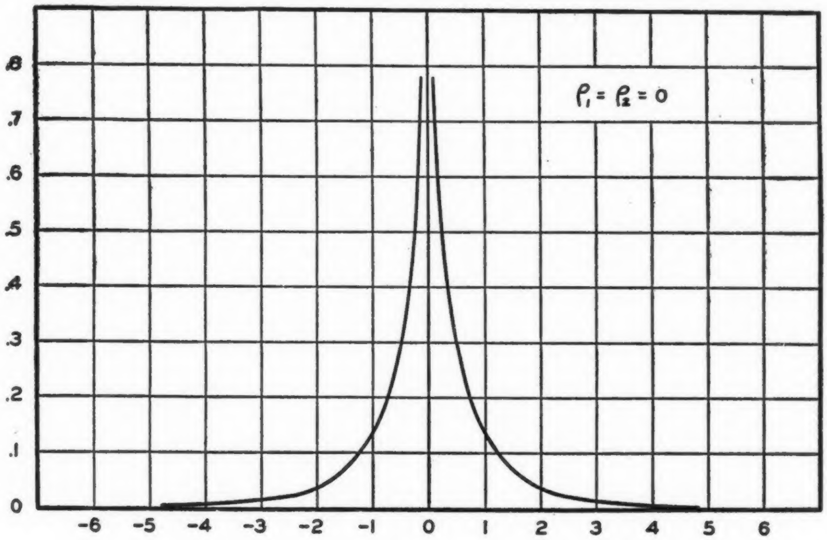



FIG. 1

TABLES OF ORDINATES OF THE DISTRIBUTION FUNCTIONS, $F(z)$ AND $F(t)$

For $\rho_1 = \rho_2 = 0, r = 0$
(Curve is symmetrical with respect to origin)
$M_z = 0, \sigma_z = 1$

| $z = t$ | $F(z) \equiv F(t)$ |
|---|---|
| 0.1 | 0.77256 |
| 0.2 | .55790 |
| 0.3 | .43887 |
| 0.4 | .35477 |
| 0.5 | .29425 |
| 0.6 | 0.24749 |
| 0.7 | .21025 |
| 0.8 | .17996 |
| 0.9 | .15493 |
| 1.0 | .13402 |
| 1.2 | 0.10138 |
| 1.4 | .07756 |
| 1.6 | .05983 |
| 1.8 | .04645 |
| 2.0 | .03625 |
| 2.4 | 0.02235 |
| 2.8 | .01395 |
| 3.2 | .00878 |
| 3.6 | .00557 |
| 4.0 | .00355 |
| 4.8 | 0.00146 |
| 5.6 | .00061 |
| 6.4 | .00026 |
| 7.2 | .00011 |
| 8.0 | .00005 |
| 9.0 | 0.00002 |
| 10.0 | .00001 |

$\rho_1 = 1, \rho_2 = 0, r = 0$
(Curve is symmetrical with respect to origin)
$M_z = 0, \sigma_z = \sqrt{2}$

| $z$ | $F(z)$ | $t$ | $F(t)$ |
|---|---|---|---|
| 0.1 | 0.58215 | 0.07 | 0.82328 |
| 0.2 | .44891 | .14 | .63485 |
| 0.3 | .37159 | .21 | .52551 |
| 0.4 | .31736 | .28 | .44882 |
| 0.5 | .27593 | .35 | .39023 |
| 0.6 | 0.24270 | 0.42 | 0.34323 |
| 0.7 | .21519 | .49 | .30432 |
| 0.8 | .19193 | .57 | .27143 |
| 0.9 | .17195 | .64 | .24318 |
| 1.0 | .15460 | .71 | .21863 |
| 1.2 | 0.12595 | 0.85 | 0.17812 |
| 1.4 | .10340 | 0.99 | .14623 |
| 1.6 | .08533 | 1.13 | .12068 |
| 1.8 | .07069 | 1.27 | .09997 |
| 2.0 | .05873 | 1.41 | .08306 |
| 2.4 | 0.04078 | 1.70 | 0.05767 |
| 2.8 | .02846 | 1.98 | .04025 |
| 3.2 | .01992 | 2.26 | .02818 |
| 3.6 | .01397 | 2.55 | .01976 |
| 4.0 | .00981 | 2.83 | .01388 |
| 4.8 | 0.00485 | 3.39 | 0.00685 |
| 5.6 | .00239 | 3.96 | .00338 |
| 6.4 | .00118 | 4.53 | .00167 |
| 7.2 | .00058 | 5.09 | .00082 |
| 8.0 | .00029 | 5.66 | .00040 |
| 9.0 | 0.00012 | 6.36 | 0.00017 |
| 10.0 | .00005 | 7.07 | .00007 |
| 11.0 | .00002 | 7.78 | .00003 |
| 12.0 | .00001 | 8.49 | .00001 |

$$\rho_1 = \rho_2 = \tfrac{1}{2},\ r = 0$$

$$M_z = 0.25,\ \sigma_z = \frac{\sqrt{6}}{2}.$$

| $z$ | $F(z)$ | $t$ | $F(t)$ |
|---|---|---|---|
| −9.6 | 0.00001 | −8.04 | 0.00001 |
| −8.8 | .00002 | −7.39 | .00002 |
| −8.0 | 0.00004 | −6.74 | 0.00005 |
| −7.2 | .00010 | −6.08 | .00012 |
| −6.4 | .00023 | −5.43 | .00028 |
| −5.6 | .00054 | −4.78 | .00066 |
| −4.8 | .00128 | −4.12 | .00157 |
| −4.0 | 0.00311 | −3.47 | 0.00381 |
| −3.6 | .00488 | −3.14 | .00598 |
| −3.2 | .00769 | −2.82 | .00942 |
| −2.8 | .01221 | −2.49 | .01495 |
| −2.4 | .01954 | −2.16 | .02393 |
| −2.0 | 0.03165 | −1.84 | 0.03876 |
| −1.6 | .05213 | −1.51 | .06384 |
| −1.2 | .08809 | −1.18 | .10788 |
| −0.8 | .15568 | −0.86 | .19066 |
| −0.4 | .30423 | −0.53 | .37259 |
| −0.2 | 0.47388 | −0.37 | 0.58036 |
| −0.1 | .64994 | −0.28 | .79598 |
| 0.1 | 0.68106 | −0.12 | 0.83409 |
| 0.2 | .51947 | −0.04 | .63619 |
| 0.4 | 0.36322 | 0.12 | 0.44484 |
| 0.8 | .21768 | .45 | .26659 |
| 1.2 | .14230 | .78 | .17427 |
| 1.6 | .09621 | 1.10 | .11783 |
| 2.0 | .06614 | 1.43 | .08100 |
| 2.4 | 0.04589 | 1.76 | 0.05620 |
| 2.8 | .03201 | 2.08 | .03920 |
| 3.2 | .02241 | 2.41 | .02745 |
| 3.6 | .01571 | 2.74 | .01924 |
| 4.0 | .01103 | 3.06 | .01351 |

$$\rho_1 = \rho_2 = \tfrac{1}{2},\ r = 0$$

$$M_z = 0.25,\ \sigma_z = \frac{\sqrt{6}}{2}.$$

| $z$ | $F(z)$ | $t$ | $F(t)$ |
|---|---|---|---|
| 4.8 | 0.00545 | 3.72 | 0.00667 |
| 5.6 | .00269 | 4.36 | .00329 |
| 6.4 | .00133 | 5.02 | .00163 |
| 7.2 | .00065 | 5.67 | .00080 |
| 8.0 | .00032 | 6.33 | .00039 |
| 8.8 | 0.00016 | 6.98 | 0.00020 |
| 9.6 | .00008 | 7.63 | .00010 |
| 10.4 | .00004 | 8.29 | .00005 |
| 11.2 | .00002 | 8.94 | .00002 |
| 12.0 | .00001 | 9.59 | .00001 |

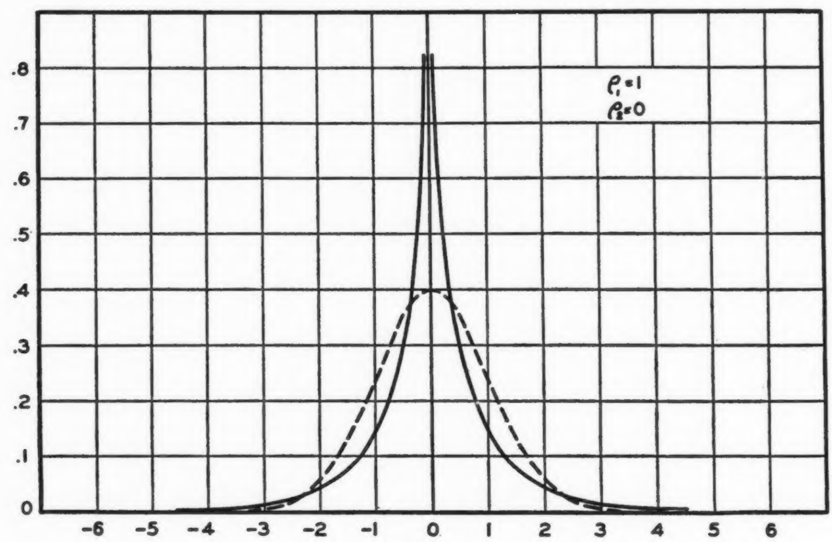



FIG. 2

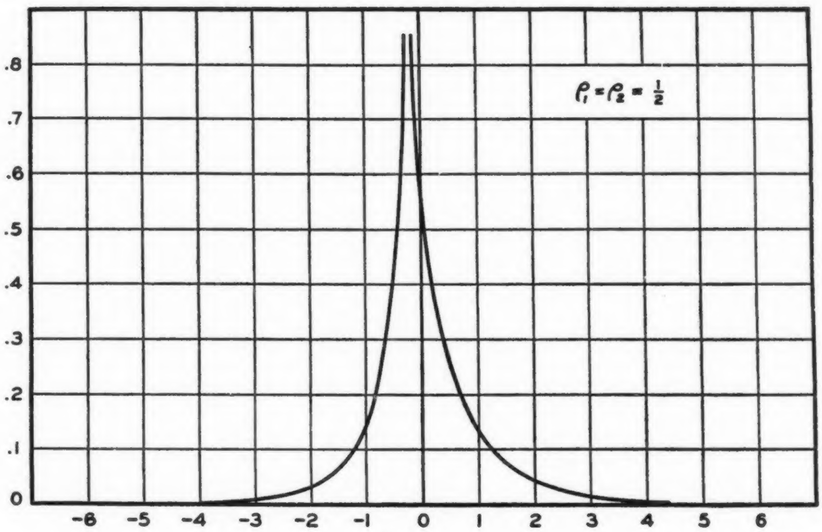



FIG. 3

UNIVERSITY OF MICHIGAN.

# A NEW EXPOSITION AND CHART FOR THE PEARSON SYSTEM OF FREQUENCY CURVES

By Cecil C. Craig

In the course of some years of teaching classes in mathematical statistics, the author has expanded the treatment of the Pearson system of frequency functions begun in the Handbook of Mathematical Statistics[1] into an exposition that he believes possesses marked advantages in unity, clarity, and elegance. This is accomplished by expressing the variable in standard units throughout and by making the two parameters $\alpha_3(\alpha_3^2 = \beta_1, \alpha_4 = \beta_2$ in Pearson's notation) and

$$\delta = \frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3}$$ [2]

fundamental in the discussion. The various formulae that arise are obtained directly and in a uniform manner and are relatively simple in form and easy to use. The criteria for the different members of the system of functions are expressed very simply in terms of $\alpha_3$ and $\delta$ and the chart corresponding to the extension of the Rhind diagram given by Pearson[3] takes on a strikingly simple form.

Following the beginning made in the Handbook, the system of Pearson frequency functions are to be found among the solutions of the differential equation

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1 t + b_2 t^2}. \tag{1}$$

For those solutions $y = f(t)$ for which,

$$(b_0 + b_1 t + b_2 t^2)\, t^n f(t)\Big]_{t=r}^{s} = 0,$$

[1] H. L. Rietz, Editor-in-Chief; Houghton-Mifflin Co., Boston (1924). See the chapter on Frequency Curves by H. C. Carver.

[2] The notation used is that of the Handbook, loc. cit., to which reference will be frequently made. The discussion of Robert Henderson, "Frequency Curves and Moments," Transactions of the Actuarial Society of America, Vol. VIII (1904), pp. 30–41, also proceeds along very similar lines, although Professor Carver was quite unaware of it when he wrote his chapter in the Handbook. The notation of the Handbook seems preferable however.

[3] Karl Pearson: Mathematical Contributions to the Theory of Evolution, XIX. Second Supplement to a Memoir on Skew Variation; Proc. Roy. Soc., A. Vol. 216 (1916), plate opposite p. 456.

if $r$ and $s$ are the extremes of the range of variation for $t$, and for which the first $n + 1$ moments over this range exist, the recursion formula for moments,

$$\alpha_n a + n\alpha_{n-1}b_0 + (n+1)\alpha_n b_1 + (n+2)\alpha_{n+1}b_2 = \alpha_{n+1}, \tag{2}$$

can be derived. Then setting $n = 0, 1, 2, 3$ we get the following expressions for the parameters, $a$, $b_0$, $b_1$, $b_2$ in terms of $\alpha_3$ and $\delta$:

$$\begin{aligned} a &= -\frac{\alpha_3}{2(1+2\delta)}, & b_1 &= \frac{\alpha_3}{2(1+2\delta)} \\ b_0 &= \frac{2+\delta}{2(1+2\delta)} & b_2 &= \frac{\delta}{2(1+2\delta)} \end{aligned} \tag{3}$$ [4]

valid except when $\delta = -\frac{1}{2}$. Below note will be taken of those solutions for which the conditions imposed in deriving (2) are not satisfied. The case in which $\delta = -\frac{1}{2}$ will be included in the discussion of the transitional types of functions.

It is useful to note that

$$-2 < \delta < 2.$$

To show this, using a well-known device, we see that

$$\int_r^s f(t)\,(t^2 + \lambda t)^2\,dt = \alpha_4 + 2\lambda\alpha_3 + \lambda^2$$

is never negative since $f(t) \geqq 0$, $r \leqq t \leqq s$, for any real $\lambda$. This requires that

$$\alpha_3^2 \leqq \alpha_4 .$$

But

$$-2 + \frac{4\alpha_4 - 3\alpha_3^2}{\alpha_4 + 3} = \delta = 2 - \frac{(\alpha_3^2 + 4)\;3}{\alpha_4 + 3}$$

and the result follows. One consequence of this is that $b_0$ cannot vanish for any Pearson frequency function possessing moments of the fourth order.

Turning now to the integration of (1) and the development of the various forms of $f(t)$ that arise, it is useful to make the preliminary statements:

1. Over the range of variation of $t$, we must have $f(t) \geqq 0$.
2. The area under curve $y = f(t)$ over the range of variation must be finite. This being true then we always determine the constant of integration so that this area is unity.
3. The range in each case is taken as the maximum one for which (1) and (2) may be secured which contains the point, $t = 0$.
4. It is sufficient throughout to take $\alpha_3 \geqq 0$ since the curve for $\alpha_3 = -k$ is only a reflection of that for $\alpha_3 = k$ through the line $t = 0$.

---

[4] See the Handbook, pp. 103, 104.

It seems best to follow the Handbook in disposing of three of the transitional types before proceeding to the main types of the system and then to the remaining transitional types.

The discussion is planned to embody a direct and uniform method of treatment, giving simple formulae for the calculation of the parameters in terms of $\alpha_3$ and $\delta$ in each case, and noting the salient features of each type of curve. The criteria for each type are expressed in terms of $\alpha_3$ and $\delta$, which for the whole system permit a simple graphical representation by means of the chart found at the end of this article. The construction of this chart is made clear in the deviation of the criteria.

**Transitional Type: The Normal Frequency Function:** $\alpha_3 = \delta = 0$

In this case (1) reduces to,

$$\frac{1}{y}\frac{dy}{dt} = -t,$$

from which

$$\text{(N)} \qquad y = c\, e^{-\frac{t^2}{2}}.$$

The range is, of course, $(-\infty, \infty)$ with $C = (2\pi)^{-\frac{1}{2}}$. On the chart, which we shall refer to as the $(\alpha_3^2, \delta)$-diagram, we see that this function corresponds to but a single point.

It may have the appearance of reasoning in a circle to use the values of the parameters given by (3), which were derived from (2), in solving (1) and then for the solution obtained examine the validity of (2). However, we may argue as follows: We will use the relations (3) as definitions of $a$, $b_0$, $b_1$, and $b_2$ in terms of $\alpha_3$ and $\delta$ which are not yet defined. Using the values of $a$ and the $b$'s given by any choice of $\alpha_3$ and $\delta$, we solve (1). If the solution is such that for it (2) may be derived, then the relations (3) are valid when $\alpha_3$ and $\delta$ have their usual meanings. For convenience let us denote the conditions for the validity of (2) by $(A)$. It is obvious that conditions $(A)$ are satisfied for

$$\text{(N)} \qquad f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

$$\textbf{Transitional types} \begin{cases} \textbf{III}, \ \alpha_3 \neq 0, & \delta = 0 \\ \textbf{X, if also } \alpha_3^2 = 4. \end{cases}$$

We get here (See the Handbook, loc. cit.):

$$\text{(III)} \qquad f(t) = \frac{A^{A^2} e^{-A^2}}{\Gamma(A^2)} (A + t)^{A^2-1} e^{-At},$$

if $A = 2/\alpha_3$, the range being $(-A, \infty)$.

It is readily verified that, since $A^2 - 1 > -1$, conditions $(A)$ are satisfied.

For $A^2 > 1$ (i.e., for $\alpha_3^2 < 4$) the curve is bell-shaped; for $A^2 < 1$ it is J-shaped with an infinite ordinate at $t = -A$. For the bell-shaped curve the mode falls at $t = -1/A$ and the mean—the mode $= 1/A = \alpha_3/2$.

For $A^2 = 1$, we have

$$f(t) = \frac{e^{-t}}{e}, \tag{X}$$

which represents a J-shaped curve with the range $(-1, \infty)$.

For $A^2 \neq 1$, the function has been designated type III, the special case as type X. On the $(\alpha_3^2, \delta)$-chart the points corresponding to type III functions fall on the line $\delta = 0$, the type X functions being represented by a single point on this line.

Turning now to the discussion of the three main types, we note that for $\delta \neq 0$, $b_2 \neq 0$ and that consequently the denominator on the right in (1) is always a quadratic which we can write in the form

$$b_2(t - r_1)(t - r_2)$$

in which neither $r_1$ nor $r_2$ can be zero (since $b_0 \neq 0$), and

$$\begin{aligned} r_1 &= \frac{-b_1 + \sqrt{b_1^2 - 4b_0b_2}}{2\,b_2} = \frac{-\alpha_3 + \sqrt{\alpha_3^2 - 4\delta(\delta + 2)}}{2\,\delta} = \frac{-\alpha_3 + \sqrt{D}}{2\,\delta}. \\ r_2 &= \frac{-\alpha_3 - \sqrt{D}}{2\,\delta}. \end{aligned} \tag{4}$$

Leaving aside the special case, $r_1 = r_2$, to be dealt with later, we can always solve (1) in the form

$$f(t) = C(t - r_1)^{m_1}(t - r_2)^{m_2} \tag{5}$$

with

$$\begin{aligned} m_1 &= \frac{a - r_1}{b_2(r_1 - r_2)} = \frac{1 + \delta}{\delta}\frac{\alpha_3}{\sqrt{D}} - \frac{1 + 2\,\delta}{\delta} \\ m_2 &= \frac{a - r_2}{b_2(r_2 - r_1)} = -\frac{1 + \delta}{\delta}\frac{\alpha_3}{\sqrt{D}} - \frac{1 + 2\,\delta}{\delta}. \end{aligned} \tag{6}$$

For $\delta < 0$, the $r$'s are real and opposite in sign; for $\delta > 0$ and $\alpha_3^2 < 4\delta(\delta + 2)$, the $r$'s are complex; and for $\delta > 0$ and $\alpha_3^2 > 4\delta(\delta + 2)$, the $r$'s are real and of the same sign. These three conditions with the additional condition that $\alpha_3 \neq 0$ give rise respectively to the *main* types of frequency functions designated I, IV, and VI. The points corresponding to them fall in simply determined areas on the $(\alpha_3^2, \delta)$-chart. The boundaries of these areas, the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta),$$

which intersects the type I and type VI areas, and the line,

$$\delta = -1/2$$

contain the points which correspond to the transitional types.

**Main Type I.** $\alpha_3 \neq 0, -1 < \delta < 0\ [\delta \neq -\frac{1}{2}, (2 + 3\delta)\alpha_3^2 \neq 4(1 + 2\delta)^2(2 + \delta)]$

For $\alpha_3 > 0$, we see that

$$r_1 < 0 < r_2 \text{ and that } |r_1| < |r_2|.$$

The range is taken to be $(r_1, r_2)$ and (5) is written

$$y = C(t - r_1)^{m_1}(r_2 - t)^{m_2}. \tag{I}$$

It is evident that the area under the curve over this interval is finite only when $m_1 + 1 > 0$ and $m_2 + 1 > 0$ and that if these inequalities hold moments of all orders exist. In this case also conditions $(A)$ are satisfied. Now

$$m_1 + 1 = -\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right)$$

$$m_2 + 1 = -\frac{1+\delta}{\delta}\left(1 + \frac{\alpha_3}{\sqrt{D}}\right),$$

and in the present case

$$1 \pm \frac{\alpha_3}{\sqrt{D}} > 0.$$

Thus $m_1 + 1$ and $m_2 + 1$ are each $> 0$ only if $\delta > -1$. On the chart, then, the points for $\delta < -1$ correspond to no frequency functions,—they fall in the "Impossible Area."

Further the type I curve will be U-shaped, J-shaped, or bell-shaped if both $m$'s are $< 0$, if the $m$'s are opposite in sign, or if both are $> 0$. We have

$$m_1 = -\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right) - 1.$$

Since for $-1 < \delta < -\frac{1}{2}$,

$$0 < -\frac{1+\delta}{\delta} < 1,$$

we see that $m_1 < 0$ $(\alpha_3 > 0)$ for $\delta$ in this interval. For $-\frac{1}{2} < \delta < 0$, $m_1 > 0$ only if

$$-\frac{1+\delta}{\delta}\left(1 - \frac{\alpha_3}{\sqrt{D}}\right) > 1,$$

which leads to the condition:

$$(2 + 3\delta)\alpha_3^2 < 4(1 + 2\delta)^2\,(2 + \delta)\,.$$

Also,

$$m_2 = -\frac{1 + \delta}{\delta}\left(1 + \frac{\alpha_3}{\sqrt{D}}\right) - 1$$

whence it is similarly seen that $m_2 > 0$ when $-\frac{1}{2} < \delta < 0$, and that generally $m_2 > 0$ only when

$$(2 + 3\delta)\alpha_3^2 < 4(1 + 2\delta)^2\,(2 + \delta)\,.$$

Thus the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2\,(2 + \delta)\,,$$

being tangent to the line $\alpha_3^2 = 0$ at $\delta = -\frac{1}{2}$, divides the type I area on the chart into three parts: Above it lie the points corresponding to U-shaped curves, to the right of it the points corresponding to J-shaped curves, and below it the points corresponding to bell-shaped curves. (Note that for $\delta < -\dfrac{2}{3}$ the curves are always U-shaped.)

Since $r_2 - r_1 > 0$ and $b_2 \gtrless 0$ accordingly as $\delta \lesseqgtr -\frac{1}{2}$, it is readily verified that $r_1 < a < r_2$ only for U- or bell-shaped curves. The sign of $a$ is always opposite to that of $\alpha_3$ for curves with a mode. Finally the constant is determined by setting

$$C\int_{r_1}^{r_2} (t - r_1)^{m_1}(r_2 - t)^{m_2}\,dt = 1\,,$$

giving

$$C = \frac{1}{\beta(m_1 + 1,\, m_2 + 1)\,(r_2 - r_1)^{m_1+m_2+1}}\,.$$

**Main Type IV:** $\alpha_3 \neq 0$, $\delta > 0$, and $\alpha_3^2 < 4\delta(\delta + 2)$

In this case we write:

$$r_1 = \frac{-\alpha_3}{2\delta} + \frac{i\sqrt{-D}}{2\delta} = -r + is\,, \qquad r_2 = -r - is\,.$$

$$m_1 = -\frac{1 + \delta}{\delta}\,\frac{\alpha_3}{\sqrt{-D}}\,i - \frac{1 + 2\delta}{\delta} = \frac{\nu i}{2} - m\,, \qquad m_2 = -\frac{\nu i}{2} - m\,.$$

With this notation (5) becomes

$$y = C[(t + r)^2 + s^2]^{-m}\left(\frac{t + r - is}{t + r + is}\right)^{\frac{\nu i}{2}},$$

and since,

$$\left(\frac{a - bi}{a + bi}\right)^{\frac{ci}{2}} = e^{c \tan^{-1} b/a} = e^{c(\pi/2 - \tan^{-1} a/b)},$$

the frequency function can be written,

$$\text{(IV)} \qquad y = Ce^{\nu\pi/2}[(t + r)^2 + s^2]^{-m} e^{-\nu \tan^{-1} \frac{t+r}{s}}.$$

It is readily seen that $m > 0$, that $\nu$ is opposite in sign to $\alpha_3$, that

$$e^{-\nu \tan^{-1} \frac{t+r}{s}}$$

can always be taken to lie between $e^{-\nu\pi/2}$ and $e^{\nu\pi/2}$, and that the range can now be taken $(-\infty, \infty)$.

In the previously discussed cases in which $\delta \leqq 0$, if the area under the curve was finite moments of all orders existed. In the present case, the area and the first four moments are always finite but this may fail to be true of moments of higher orders. For, since $0 < \delta < 2$,

$$m = \frac{1 + 2\delta}{\delta} > \frac{5}{2},$$

and the integral,

$$\int_{-\infty}^{\infty} t^n f(t)\, dt$$

for $f(t)$ given by (IV) will be finite for $n \leqq 4$ and infinite for $n = 5$ if $\delta \geqq 1$. In order for the $n$-th moment to exist we must have

$$2m > n + 1$$

or

$$\delta < \frac{2}{n - 3}.$$

Pearson designated as *heterotypic* those members of his system of frequency functions for which the eighth moment failed to exist. (In such a case the standard deviation of the fourth moment in samples would be infinite.) Setting $n = 8$, we get $\delta = 2/5$ as the deadline on the $(\alpha_3^2, \delta)$-chart.

It was apparent that conditions ($A$) were satisfied for $-1 < \delta < 0$. (It will appear below that the case in which $\delta = -\frac{1}{2}$ is no exception.) For $\delta > 0$ it will be seen that it is generally true, as in the present case, that the formulae (2) and (3) can be derived if $\alpha_{n+2}$ exists, i.e., if

$$\delta < \frac{2}{n - 1}.$$

To determine $C$, on setting the integral of (V) over the interval $(-\infty, \infty)$ equal to unity, we get

$$C = \frac{s^{2m-1}}{G(2m-2, \nu)}$$

in which

$$G(2m-2, \nu) = \int_0^{\pi} \sin^{2m-2} \varphi e^{\nu\varphi} \, d\varphi \text{ [5]} \qquad \left(\varphi = \frac{\pi}{2} - \tan^{-1} \frac{t+r}{s}\right).$$

**Main Type VI:** $\alpha_3 \neq 0, \delta > 0, \alpha_3^2 > 4\delta(\delta + 2)\,[(2 + 3\delta)\alpha_3^2 \neq 4(1 + 2\delta)^2(2 + \delta)]$

The conditions specify the remaining area on the chart. This may be left in the form

$$y = C(t - r_1)^{m_1}(t - r_2)^{m_2}. \tag{5}$$

Now $r_1$ and $r_2$ are both opposite in sign to $\alpha_3$, which, as usual, we will consider positive, and $|r_2| > |r_1|$. Always $m_2 < 0$ and $m_1 \gtrless 0$ accordingly as

$$(2 + 3\delta)\alpha_3^2 \lessgtr 4(1 + 2\delta)^2\,(2 + \delta).$$

We note that

$$a - r_2 = b_2(r_2 - r_1)m_2 > 0,$$

since now $b_2 > 0$, and that

$$a - r_1 = b_2(r_1 - r_2)\,m_1$$

has the same sign as $m_1$. Finally $a < 0$.

Thus for $\alpha_3 > 0$ and $m_1 > 0$, the point $t = a$ on the axis of $t$ lies to the right of both $t = r_1$ and $t = r_2$. Also

$$m_1 + m_2 = -\frac{2(1 + 2\delta)}{\delta} = -4 - \frac{2}{\delta}.$$

The range is taken $(r_1, \infty)$, the curve being bell-shaped when $m_1 > 0$. If $m_1 < 0$, the curve is J-shaped, $t = a$ now lying to the left of $t = r_1$.

Since

$$m_1 + m_2 < -5, \text{ and } m_1 + 1 > 0,$$

the area and the first four moments always exist. In order for the $n$-th moment to be finite, we must have

$$-(m_1 + m_2) > n + 1$$

which is the same condition as in the case of the type IV function, giving the same deadline, $\delta = 2/5$.

[5] Cf: Tables for Statisticians and Biometricians, Cambridge Univ. Press, Part I, 2nd edition (1924), p. lxxxi.

If the origin be shifted to the point, $t = r_2$, we have writing,

$$t - r_2 = z, \qquad r_1 - r_2 = \alpha,$$

for the type VI function the expression,

$$\text{(VI)} \qquad y = Cz^{m_2}(z - \alpha)^{m_1},$$

with the range $(\alpha, \infty)$. Finally

$$C = \frac{1}{\alpha^{m_1+m_2+1}\beta(m_1 + 1, -m_1 - m_2 - 1)}.$$

**Transitional Type II:** $\alpha_3 = 0, -1 < \delta < 0.$ $(\delta \neq -\frac{1}{2})$

In this case,

$$r_1 = -r_2 = \frac{\sqrt{D}}{\delta} < 0$$

$$m_1 = m_2 = -\frac{1 + 2\delta}{\delta} \gtrless 0 \quad \text{accordingly as } \delta \gtrless -\tfrac{1}{2}.$$

The frequency function is a special case of type I; setting,

$$-r_1 = r_2 = S$$

$$m_1 = m_2 = M,$$

we can write it in the form,

$$\text{(II)} \qquad y = C(S^2 - t^2)^M.$$

As in all cases in which $\alpha_3 = 0$, the curve is symmetrical about the mean.[6] As in the type I case, the area and moments do not exist for $\delta \leqq -1$; for $-1 < \delta < -\frac{1}{2}$, the curve is U-shaped; for $-\frac{1}{2} < \delta < 0$, it is bell-shaped. The range is, of course, $(-S, S)$.

Finally,

$$C = \frac{1}{(2S)^{2M+1}\beta(M + 1, M + 1)}.$$

**Transitional Type VII;** $\alpha_3 = 0, \delta > 0$

This function may be regarded as a special case of type IV, with

$$r = 0, \quad s = \frac{\sqrt{4\delta(\delta + 2)}}{2\delta} > 0, \quad \nu = 0, \quad \text{and} \quad m = \frac{1 + 2\delta}{\delta} > 0,$$

---

[6] It follows at once from the recursion formula,

$$\alpha_{n+1} = \frac{n}{2 - (n - 2)\delta}[(2 + \delta)\alpha_{n-1} + \alpha_3 \alpha_n],$$

obtained from setting the expressions (3) in (2), that on changing the sign of $\alpha_3$, the signs of all the odd moments are changed.

and we write the function:

$$\text{(VII)} \qquad y = C(t^2 + s^2)^{-m}.$$

The type VII function may equally well be derived from the type II function by noting that

$$S = is \text{ and } M = -m.$$

The range is $(-\infty, \infty)$ however and for $\delta \geqq 2/5$ the function is heterotypic. Finally

$$C = \frac{s^{2m-1}}{\sqrt{2\pi}} \frac{\Gamma(m)}{\Gamma\left(\frac{2m-1}{2}\right)}.$$

**Transitional Type V; $\alpha_3 \neq 0, \delta > 0, \alpha_3^2 = 4\delta(\delta + 2)$**

Here

$$r_1 = r_2 = -r$$

and we return to (1) to derive the form of the function, writing it: (The type V can also be derived as a limiting form of type VI)

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_2(t + r)^2}.$$

On integration we get

$$y = C(t + r)^{-\frac{1}{b_2}} e^{-\frac{a+r}{b_2(t+r)}}$$

$$= C(t + r)^{-\frac{2(1+2\delta)}{\delta}} e^{-\frac{\alpha_3(1+\delta)}{\delta^2(t+r)}}.$$

$$\text{(V)} \qquad = C(t + r)^{-2m} e^{-\frac{2r(m-1)}{t+r}}.$$

We note that $r$ has the same sign as $\alpha_3$ and that $m = 2 + 1/\delta$. The range is taken to be $(-r, \pm \infty)$ accordingly as $\alpha_3 \gtrless 0$. The curve is always bell-shaped. In order for the $n$-th moment to exist we must have as always when $\delta > 0$,

$$4 + 2/\delta > n + 1$$

leading to the same conclusions as in the type IV or VI case. Finally

$$C = \frac{[2r(m - 1)]^{2m-1}}{\Gamma(2m - 1)}.$$

**Transitional Type VIII; $\alpha_3 = 0, \delta < -\frac{1}{2}, (2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$**

The function is a special case of type I in which $m_1 < 0$ and $m_2 = 0$. But when $m_2 = 0$, $m_1 = -2m$, and the frequency function becomes

$$\text{(VIII)} \qquad y = C(t - r_1)^{-2m}.$$

The range is $(r_1, r_2)$, the curve being J-shaped with an infinite ordinate at $t = r_1$ and a finite one at $t = r_2$. In this case,

$$C = \frac{1 - 2m}{(r_2 - r_1)^{1-2m}} \,. \qquad (1 - 2m > 1)$$

**Transitional Type IX:** $\alpha_3 \neq 0, -\frac{1}{2} < \delta < 0, (2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$

We have another special type I function in which $m_1 = 0$ and $m_2 = -2m > 0$. The function is

$$\text{(IX)} \qquad y = C(r_2 - t)^{-2m}$$

the range still being $(r_1, r_2)$, the curve being J-shaped with a finite ordinate at $t = r_2$. $C$ has the same value as in the type VIII case.

**Transitional Type XI;** $\alpha_3 \neq 0, 0 < \delta < 2/5, (2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2(2 + \delta)$

The function is a special type VI in which $m_1 = 0$, and $m_2 = -2m < 0$, and we may write it

$$\text{(XI)} \qquad y = C(t - r_2)^{-2m}$$

with the range still $(r_1, \infty)$. The curve is J-shaped with a finite ordinate at $t = r_1$. Again,

$$C = \frac{2m - 1}{(r_1 - r_2)^{2m-1}} \qquad \left(2m - 1 = 3 + \frac{2}{\delta}\right)$$

**Transitional Type XII:** $\delta = -\frac{1}{2}$ .

If $\delta = -\frac{1}{2}$, the four linear equations derived from (2) from which the values of $a$, $b_0$, $b_1$, and $b_2$ in (3) are derived are inconsistent. We can however set the values (3) in the differential equation (1) and from its limiting form as $\delta \rightarrow -\frac{1}{2}$, derive the function appropriate to this case.

We obtain

$$\frac{1}{y}\frac{dy}{dt} = \frac{-\alpha_3 - 2(1 + 2\delta)t}{(2 + \delta) + \alpha_3 t + \delta t^2}$$

and if $\delta = -\frac{1}{2}$, this becomes

$$\frac{1}{y}\frac{dy}{dt} = \frac{2\,\alpha_3}{t^2 - 2\,\alpha_3 t - 3} = \frac{2\,\alpha_3}{(t - r_1)\,(t - r_2)}$$

with

$$r_1 = \alpha_3 - \sqrt{\alpha_3^2 + 3}\,, \qquad r_2 = \alpha_3 + \sqrt{\alpha_3^2 + 3}\,.$$

On integration,

$$y = C'(t - r_1)^{m_1}\,(t - r_2)^{m_2}\,,$$

in which

$$m_1 = -\frac{\alpha_3}{\sqrt{\alpha_3^2 + 3}}, \qquad m_2 = \frac{\alpha_3}{\sqrt{\alpha_3^2 + 3}}.$$

We observe that ($\alpha_3 > 0$)

$$r_2 > 0 > r_1, \qquad |r_2| > |r_1|$$

$$m_2 = -m_1 > 0.$$

Taking the range to be $(r_1, r_2)$, we write,

(XII) $$y = C\left(\frac{r_2 - t}{t - r_1}\right)^{m_2},$$

the curve being J-shaped. Here

$$C = \frac{1}{(r_2 - r_1)\,\beta(1 - m_2, 1 + m_2)}.$$

The values of the parameters and the form of the function can also be derived as a special type I function in which $\delta = -\frac{1}{2}$.

Finally we note that for $\alpha_3 = 0$, (XII) reduces to

$$y = C$$

thus including the rectangular distribution function among the Pearson system.

In the course of the above discussion a system of criteria for the various types of functions has been set up in terms of $\alpha_3$ and $\delta$, in terms of which in every case the parameters may be readily calculated. The $(\alpha_3^2, \delta)$-chart which makes these criteria visual is comparatively simple to construct and is strikingly simple in appearance. Besides the lines,

$$\delta = -1, \qquad \delta = -\frac{1}{2}, \qquad \delta = 0, \qquad \delta = \frac{2}{5}, \quad \text{and} \quad \alpha_3 = 0,$$

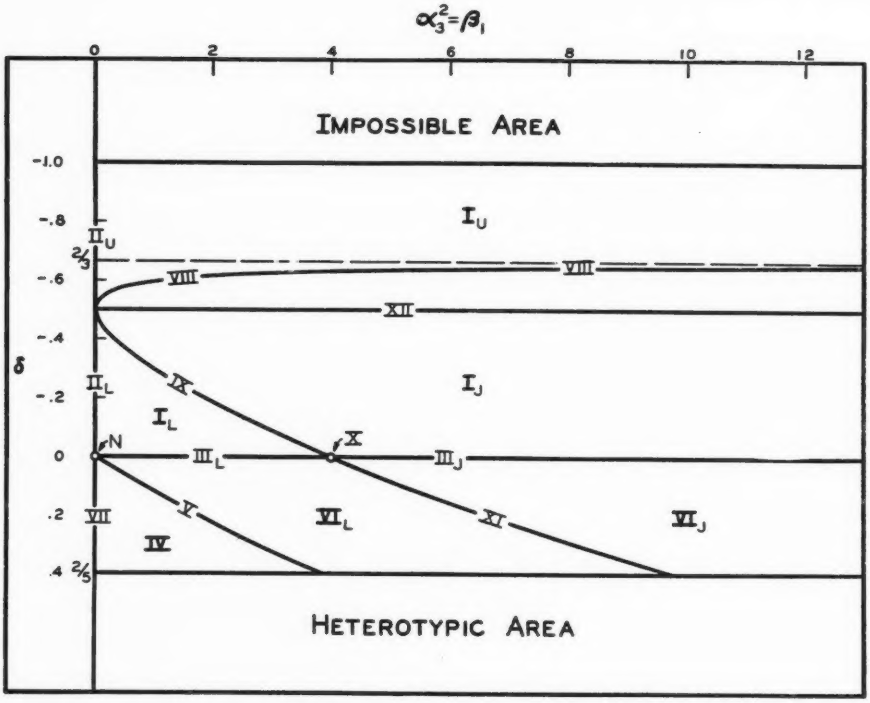
it contains only the curves

$$\alpha_3^2 = 4\delta(\delta + 2)$$

on which the points corresponding to the type V function lie, and the curve,

$$(2 + 3\delta)\alpha_3^2 = 4(1 + 2\delta)^2\,(2 + \delta)$$

on which the points corresponding to the functions of types VIII, IX, X, and XI are found. I must take occasion to express my thanks to Mr. Simon Yang who constructed this chart for me.




THE ($\alpha^2$, $\delta$) CHART FOR THE PEARSON SYSTEM OF FREQUENCY CURVES
(The subscript $L$ refers to bell-shaped curves)

THE UNIVERSITY OF MICHIGAN.

# RANK CORRELATION AND TESTS OF SIGNIFICANCE INVOLVING NO ASSUMPTION OF NORMALITY*, †

By Harold Hotelling and Margaret Richards Pabst

## 1. Dependence of Tests of Significance on Normality

The powerful tests of significance, largely the work of R. A. Fisher, which have been revolutionizing statistical theory and practice, are in the main based on the assumption of a normal distribution in a hypothetical population from which the observations are a random sample. The nature and extent of the errors likely to result from the application of a test of significance assuming normality, where normality does not really exist, have been the subject of investigations both experimental and mathematical,[1] which however have not produced satisfactory substitutes for Fisher's methods. A false assumption of normality does not usually give rise to serious errors in the interpretation of simple means, since the distribution of a mean of any considerable number of cases is very nearly normal, no matter what the nature of the parent population, so long as it does not fall within a certain class having infinite range, and including the Cauchy distribution. The sampling distributions of second-order statistics are however more seriously disturbed by lack of normality, as is evident from their standard errors. For example the variance $(\mu_4 - \mu_2^2)/n$ of sample variances is much affected if $\mu_4/\mu_2^2$ differs considerably, as it often does, from the value 3 which it takes for a normal distribution. Likewise the approximate variance of the correlation coefficient,

$$\sigma_r^2 = \frac{1}{n\mu_{20}\mu_{02}}\left\{\mu_{22} + \frac{\mu_{40}\mu_{11}^2}{4\mu_{20}^2} + \frac{\mu_{04}\mu_{11}^2}{4\mu_{02}^2} - \frac{\mu_{31}\mu_{11}}{\mu_{20}} - \frac{\mu_{13}\mu_{11}}{\mu_{02}} + \frac{\mu_{22}\mu_{11}^2}{2\mu_{20}\mu_{02}}\right\},$$

* Research under a grant-in-aid from the Carnegie Corporation of New York.

† Presented to the American Mathematical Society at New York, Oct. 26, 1935.

[1] J. L. Carlson, *A Study of the Distribution of Means Estimated from Small Samples by the Method of Maximum Likelihood for Pearson's Type II Curve*, Unpublished M. A. Thesis, Leland Stanford Junior University, 1931.

Leone Chesire, Elena Oldis and Egon S. Pearson, *Further Experiments on the Sampling Distribution of the Correlation Coefficient*, Journal of the American Statistical Association, June, 1932, pp. 121–128.

Victor Perlo, *On the Distribution of Student's Ratio for Samples of Three Drawn from a Rectangular Distribution*, Biometrika, Vol. XXV, Parts I and II, May, 1933, pp. 203–204.

Paul R. Rider, *On the Distribution of the Ratio of Mean to Standard Deviation in Small Samples from Non-Normal Universes*, Biometrika, Vol. XXI, Parts I to IV, December, 1929, pp. 124–143.

H. L. Rietz, *Note on the Distribution of the Standard Deviation, etc.*, Biometrika, Vol. XXIII, 1931, pp. 424–426.

W. A. Shewhart and F. W. Winters, *Small Samples—New Experimental Results*, Bell Telephone Laboratories, Reprint B-327, July, 1928.

where $\mu_{ij}$ is the mean value of $x^i y^j$, and $\mu_{10} = \mu_{01} = 0$, may be substantially different from the value $(1 - \rho^2)^2/n$ commonly used, to which it reduces if the population has the bivariate normal distribution. It is however remarkable that if the variates are really independent, so that $\mu_{11} = 0$ and $\mu_{22} = \mu_{20}\mu_{02}$, this formula reduces to

$$\sigma_r^2 = \frac{1}{n}, \tag{1}$$

regardless of the form of the distribution. It should of course be remembered that these formulae give only the first term of an expansion in inverse powers of $n$, and also that the standard error fails for small samples to characterize the distribution adequately. But the sensitiveness of the standard error formula to deviations from normality in the population is a symptom of the grave dangers in using even those distributions which for normal populations are accurate, in the absence of definite evidence of normality.

To substitute in standard error formulae values of the higher moments estimated from the data does not meet the difficulty satisfactorily, since these higher moments are themselves subject to sampling errors which are often large, and since no exact distributions can ever be obtained in this way. The use of an arbitrary system of distributions such as the Pearson curves is subject to the same criticisms as that of the normal distribution. These and other special distributions may indeed be justified in special cases by general reasoning; an example of this in introducing a measure of relationship other than the correlation coefficient is to be found in the genetic discussion of Chapter 9 of Fisher's "Statistical Methods for Research Workers." But for a great deal of statistical work no such a priori reasoning is available and sufficient to specify a distribution in sufficient detail. If a specific form of distribution other than the normal can be relied on in a particular case, the mathematical problem of finding the exact distribution of the appropriate statistic will still commonly be found difficult or impossible.

## 2. Tests Independent of Normality Assumptions

A set of problems is thus encountered regarding the nature and methods of statistical inference possible without assuming any particular distribution of the variates in the population from which we have a sample. Tests of significance underlying such inferences must clearly be invariant under all transformations of each variate. We are thus forced to rely for our information on relations of *order*, or of qualitative classification, rather than upon magnitudes, excepting insofar as we can use inequalities such as that of Tchebycheff. Classification leads to the use of contingency tables, from which accurate probabilities are calculable for testing whether or not the two or more principles of cross-classification used are independent. If the probability obtained is so small as to render it incredible that independence exists, the further problem arises of measuring the degree of relationship; but in the absence of special assumptions, such as that

of the bivariate normal distribution, or those in Fisher's genetic example mentioned above, the problem of measuring degree of relationship is insoluble. Any measure of degree of relationship will change its value, unless this value corresponds to independence, when transformations other than those of a restricted class are applied to one of the variates. The problem of measuring *degree* of relationship, or correlation, is thus of quite a different character from that of testing the *existence* of a relationship, which is equivalent to absence of independence. The existence of correlation may be detected by methods of rank order or of classification; these can never, by themselves, be sufficient for its measurement.

To test the deviation of the center of a symmetrical population from some definite hypothetical value, Student's distribution, which is appropriate when the population is normal, may be replaced by the binomial distribution, which will sometimes show that the preponderance of cases on one side of the hypothetical value is too great to admit the hypothesis. Fisher applied this principle to Student's original example, showing at the same time that it can in certain cases be used to test the significance of the difference between the means of two samples.[2] Both this type of test and the use of contingency tables with grouped values of variates bring out clearly the fact that abandonment of the assumption of normality is equivalent to a certain loss of information, larger samples being required to make up for the lack of knowledge of the form of the population. The loss of information is greater for contingency tables arranged according to the values of the variates than when an appropriate method of rank correlation is used, for the contingency table may be regarded as derived from the ranks by grouping them, thus discarding some of the information.

We shall in §8 illustrate a combination of rank and contingency methods suitable for utilizing simultaneously two kinds of information contained in grouped data.

For large samples a method of treatment for which a great deal is to be said in many cases consists of replacing the observed variate by a new variate $x$ to which a value is assigned for each individual or frequency class by interpolation in a table of the normal probability integral, in such a way that the distribution of $x$ in the sample approximates normality. If this is done for each of two variates which do not have the bivariate normal distribution, the transformed values $x$ and $y$ may also lack the bivariate normal distribution, even approximately, though each is normally distributed, so far as we can speak of a sample as being normally distributed. Even if the bivariate distribution is normal, the correlation coefficient of $x$ and $y$ will not have the same distribution as the correlation coefficient in samples drawn from a bivariate normal distribution, since in the latter case the distributions of $x$ and $y$ separately would in most samples be less nearly normal than when the transformation to approximate normality is applied. From these considerations it follows that for the detection of correlation the normalizing transformation cannot be said in general to be the best

[2] R. A. Fisher, *Statistical Methods for Research Workers*, Art. 24, end.

method, even for large samples, though it may be a useful preliminary to the application of the method of least squares or to the use of correlation coefficients significantly different from zero in certain cases.

### 3. The Rank Correlation Coefficient

Suppose that $n$ individuals are arranged in two orders with respect to two different attributes. Thus we might arrange a freshman class in order according to their grades in a language examination, and also according to their mathematical grades. As another example, we might be able to obtain ratings of various states with respect to penal law or practice, and also with respect to amount of crime. Continuous variates expressing these qualities are likely not to be normally distributed, so that the product-moment correlation coefficient $r$ cannot be expected to have the exact distribution known for it in the case of samples from a normal population. We may therefore resort to the ranks, ignoring any exact values that have been assigned.

Calling $X_i$ the rank of the $i$th individual with respect to one attribute, and $Y_i$ his rank with respect to the other, so that $(X_1, X_2, \cdots, X_n)$ and $(Y_1, Y_2, \cdots, Y_n)$ are two permutations of the numbers $(1, 2, \cdots, n)$, let us put $x_i = X_i - \bar{x}$, $y_i = Y_i - \bar{y}$, where

$$\bar{x} = \bar{y} = \frac{n+1}{2}$$

The rank correlation coefficient is defined as

$$r' = \frac{\Sigma\, xy}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}}, \tag{2}$$

the sums being over the $n$ values in the sample. Now since the sum of the first $n$ integers is $n(n+1)/2$, and the sum of their squares is $n(n+1)(2n+1)/6$, we have

$$\begin{aligned} \Sigma\, x^2 &= \Sigma\,(X - \bar{x})^2 = \Sigma\, X^2 - (\Sigma\, X)^2/n \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n^3 - n}{12}, \end{aligned} \tag{3}$$

and $\Sigma y^2$ has the same value. Also, if we put $d_i$ for the difference between the two ranks for the $i$th individual, so that

$$d_i = X_i - Y_i = x_i - y_i,$$

we have

$$\Sigma\, d^2 = \Sigma\, x^2 - 2\,\Sigma\, xy + \Sigma\, y^2 = \frac{n^3 - n}{6} - 2\,\Sigma\, xy\,.$$

Substituting in (2) the value of $\Sigma xy$ found from this equation, and also the values just obtained for $\Sigma x^2$ and $\Sigma y^2$, we have:

$$r' = 1 - \frac{6\,\Sigma\, d^2}{n^3 - n}\,. \tag{4}$$

This is the most convenient formula for computing $r'$.

Compared with certain other tests of correlation based on order, such as $\Sigma \mid d \mid$, or the number of inversions required to pass from one permutation of the $n$ numbers to the other, $r'$ appears to be a sensitive index of relationship, since for a given value of $n$ it possesses a greater number of distinct values. But to assert without qualification that $r'$ or any other statistic is the best possible test of correlation based on order relations alone would be meaningless. Indeed, a particular type of bivariate distribution might well have a parameter representing correlation whose significance could best be detected by a test adapted only to this particular bivariate distribution. However the rank correlation coefficient has properties that point to its value in more general use than it has heretofore received. It has been regarded chiefly as a more easily calculable substitute for the product-moment coefficient $r$. Karl Pearson has remarked that the rank correlation coefficient is the easier to compute for samples smaller than approximately forty, while $r$ involves less labor for larger samples.

The great value of the rank correlation coefficient appears to us to consist in its use as a test of the existence of correlation, a test capable of exact interpretation in terms of probability, without any assumption of a normal or other special bivariate distribution. If a bivariate distribution is specified by $f(x, y)\, dx\, dy$, the condition of independence is that $f(x, y)$ shall be the product of a function of $x$ by a function of $y$. If we put

$$\xi = \int_{-\infty}^{\infty} \int_{0}^{x} f(x', y')\, dx'\, dy'\,, \qquad \eta = \int_{0}^{y} \int_{-\infty}^{\infty} f(x', y')\, dx'\, dy'\,, \tag{5}$$

using the inner integral sign in each case to correspond to the inner differential, then each of the quantities $\xi$ and $\eta$ is distributed with uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$; and if $x$ and $y$ are independent, then $\xi$ and $\eta$ are also independent. The correlation $\rho'$ of $\xi$ with $\eta$ may be called the rank correlation of $x$ and $y$ in the population. It will vanish in case of independence. It is for this case that we shall obtain in §§5, 6 and 7 the exact probability test for $r'$ in small samples, the exact standard error and fourth moment, and asymptotic values for the higher moments, with a demonstration that, for sufficiently large samples, $r'$ can be treated as normally distributed. In §9 we shall present, in a revised and simplified form, certain work of Karl Pearson relative to the estimation of the correlation $\rho$ in a bivariate normal distribution, and apply the results to discuss the question of the importance of the lost information when measurements are replaced by ranks.

## 4. History of Rank Correlation Theory

Rank correlation seems to have had its origin in the method of representing the distribution of a variate by grades or percentiles introduced by Francis

Galton.[3] Later Spearman[4] proposed that rank be considered in place of the variate, and suggested that the correlation of ranks be used as a measure of the degree of dependence of the variates. Spearman also introduced the "footrule of correlation" based on $\Sigma \mid d \mid$.

The principal memoir on rank correlation is by Karl Pearson.[5] Assuming an underlying normal distribution, Pearson obtains a relation equivalent to

$$\rho = 2 \sin \frac{\pi}{6} \rho' , \tag{6}$$

where $\rho$ is the correlation of $x$ and $y$ in the population, and $\rho'$ is the correlation of uniformized variates $\xi$ and $\eta$ defined by (4). An estimate $r''$ of $\rho$ may be based on the rank correlation $r'$, in accordance with (6), by writing

$$r'' = 2 \sin \frac{\pi}{6} r' . \tag{7}$$

Pearson finds the first few terms of infinite series giving the standard errors of $r'$ and $r''$. He deals similarly with the estimation of correlation by means of $\Sigma \mid d \mid$. The paper contains a neat proof, attributed to Student, of the probable error of $r'$ under conditions of independence. It was this proof that suggested the analysis of §§6 and 7 below. This long memoir is very difficult to read and interpret accurately, owing chiefly to the failure to distinguish clearly between sample and population.

The use of the probable error formulae is valid only if the distributions of $r'$ and $r''$ are sensibly normal. The question of approximate normality thus raised is investigated for the first time in the present paper. In order to use these formulae it is necessary to assume not only (1) that the underlying population has the bivariate normal distribution (an assumption which requires more than that each variate be normally distributed), (2) that the first few terms of the infinite series are enough, and (3) that the distributions of $r'$ and $r''$ are practically normal, but also (4) that sample values can be put for population values in the formulae, or that population values are known independently or can be assumed. It is probably this last condition that has been least understood and has led to the greatest number of false conclusions regarding the significance of data.

A note by W. C. Eells[6] presents a compilation of numerous textbook versions of the probable errors of $r'$ and $r''$, all differing from each other and from Pear-

---

[3] Francis Galton, *Natural Inheritance*, Macmillan, 1889, Chaps. 4 and 5.

[4] C. Spearman, *The Proof and Measurement of Association Between Two Things*, American Journal of Psychology, Vol. 15, 1904.

[5] Karl Pearson, *On Further Methods of Determining Correlation*, Drapers' Company Research Memoirs, Biometric Series IV, Mathematical Contributions to the Theory of Evolution, XVI, London, Dulau, 1907.

[6] W. C. Eells, *Formulas for Probable Errors of Coefficients of Correlation*, Journal of the American Statistical Association, Vol. 24, 1929, p. 170.

son's. Taking Pearson's formulae as correct, without discussing the assumptions implicit in their use, Eells presents a table for calculating the probable errors of $r$, $r'$ and $r''$.

### 5. Significance of Rank Correlation in Small Samples

If the variates are independent we may without loss of generality assign the values $1, 2, \cdots, n$ in order to $X_1, X_2, \cdots, X_n$, and regard the $Y$'s as made up by any one of the $n!$ permutations of these numbers, all permutations being equally probable. The probability of any particular value of $r'$ is thus proportional to the number of permutations giving rise to this value. These may be enumerated with the help of (4). Thus for $n = 2$, each of the values $\pm 1$ has the probability $\frac{1}{2}$. For $n = 3$, the possible values of $r'$ are $-1, -\frac{1}{2}, \frac{1}{2}, 1$, with respective probabilities 1/6, 1/3, 1/3, 1/6. For $n = 4$ the values 1, 4/5, 3/5, 2/5, 1/5, 0 have the respective probabilities 1/24, 1/8, 1/24, 1/6, 1/12, 1/12.

From (2) it is evident that the distribution of $r'$ in case of independence is symmetrical, since each permutation is exactly as probable as that of directly opposite order, and since a change of sign of all the $x$'s or $y$'s changes the sign of $r'$ without affecting its absolute value. It is clear also that the values $r' = \pm 1$, corresponding to the two variates being in the same or opposite orders, are the extreme ones, and have each a probability $1/n!$. The next greatest value of $|r'|$ corresponds to the interchange of two consecutive individuals, who may be selected in $n - 1$ ways and makes $\Sigma d^2 = 2$. Thus the values $\pm(1 - 12/[n^3 - n])$ occur with probability $(n - 1)/n!$ each. Next to these, corresponding to $\Sigma d^2 = 4$, are the values $\pm(1 - 24/[n^3 - n])$, whose probabilities are each $(n - 2)(n - 3)/2(n!)$, since the numbers of pairs of mutually exclusive consecutive pairs in a sequence of $n$ is $(n - 2)(n - 3)/2$. In like manner, but with greater complexity, it appears that the probability of the value $1 - 36/[n^3 - n]$ is $\dfrac{(n - 3)(n - 4)(n - 5) + 12(n - 2)}{6(n!)}$. Easy calculation from these results shows that, if we require for significance a probability $P = .01$ of a value of $|r'|$ as great as or greater than the value observed, then for samples of 5 it is impossible to obtain a significant value; for $n = 6$, significance requires that $r' = \pm 1$; and for $n = 7$ the significant values of $|r'|$ are 25/28 and more. For the less stringent standard $P = .05$, a unit correlation only is significant in a sample of 5; while 29/35 is not, but 31/35 is, significant in a sample of 6.

### 6. The Standard Error and Fourth Moment

For large samples the exact calculation of probabilities becomes very laborious, and we are forced to resort to approximations. The first step in the available approximations is the determination of the standard deviation of the distribution. The square of this quantity, the second moment or variance of $r'$, may, since the mean value of $r'$ in case of independence is zero, be written

$$\sigma_{r'}^2 = \mu_2 = Er'^2,$$

the symbol $E$ denoting the expectation or mean value of the quantity following. The operation $E$ has the properties that the expectation of a sum is the sum of the expectations of the terms, the expectation of the product of *independent* variates is the product of their expectations, and the expectation of the product of a constant by a variate is the product of the constant by the expectation of the variate. It is particularly to be noted that the first of these properties holds whether the terms of the sum are mutually independent or not.

From (2) and (3) we have

$$r' = \frac{12\,\Sigma\, xy}{n^3 - n}. \tag{8}$$

Now we may regard $x_1, x_2, \cdots, x_n$ as taking the same values in all samples, these values being centered at zero and differing consecutively by unity. The $y$'s are then variates, not independent of each other, taking this same set of values, but in a manner varying from sample to sample by chance. For any particular $y$, for example that associated with $x_1$, the chance distribution has moments of the form

$$Ey^p = \frac{\Sigma\, x^p}{n} = \frac{\Sigma\, y^p}{n} = \frac{s_p}{n}, \tag{9}$$

if we denote by $s_p$ the sum of the $p$th powers of the $n$ numbers differing consecutively by unity and centered at zero. It is clear that, for every odd value of $p$, $s_p = 0$. Also, from (3),

$$s_2 = \frac{n^3 - n}{12}.$$

In view of these facts, we have from (8),

$$\sigma^2_{r'} = Er'^2 = \frac{E(\Sigma\, xy)^2}{s_2^2} = \frac{\Sigma\, x^2\, Ey^2 + 2\,\Sigma\, x_1x_2 E y_1 y_2}{s_2^2},$$

where $\Sigma x_1x_2$ stands for the sum of all the $n(n-1)/2$ *different* terms obtained by permuting the subscripts. We have

$$E\, y_1 y_2 = \frac{2\,\Sigma\, x_1x_2}{n(n-1)};$$

also

$$2\,\Sigma\, x_1x_2 = s_1^2 - s_2 = -s_2.$$

Combining these results we have:

$$\sigma^2_{r'} = \frac{1}{s_2^2}\left\{\frac{s_2^2}{n} + \frac{s_2^2}{n(n-1)}\right\} = \frac{1}{n-1}. \tag{10}$$

This is the formula obtained by Student and incorporated in Pearson's memoir.

Any desired moment of $r'$ may be obtained in this manner. However the complexity of the calculation increases rapidly with the order of the moment, and the derivation of even the fourth moment is too long to be included in this paper. The value obtained for the fourth moment is

$$\mu_4 = \frac{3(25n^4 - 13n^3 - 73n^2 + 37n + 72)}{25n(n+1)^2(n-1)^3}.$$

It will be observed immediately that the kurtosis, $\beta_2 = \mu_4/\mu_2^2$, approaches the normal value 3 as $n$ increases.

For values of $n$ which are not small enough for the exact probabilities to be computed easily, the Tchebycheff inequality,

$$P \leqq \frac{1}{(n-1)\, r'^2}, \tag{11}$$

where $P$ is the probability of a deviation exceeding $r'$, will often be of service. Thus, if $n = 25$ and $r' = .9$, (11) shows that $P$ is less than .05, so that the evidence for existence of a relationship should by an ordinary standard be regarded as significant. However this does not in general give an accurate approximation to $P$, nor do the similar inequalities involving the higher moments.

### 7. The Higher Moments and the Approach to Normality

A general moment of $r'$ of even order is defined by

$$\mu_{2\alpha} = E\, r'^{2\alpha} = \frac{1}{s_2^{2\alpha}} E\,(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}. \tag{12}$$

When the parenthesis is expanded we may take the expectation term by term, regarding the $x$'s as constants. Now

$$E\, y_1^{2\alpha} = \frac{\Sigma\, x_1^{2\alpha}}{n}, \qquad E\, y_1^{2\alpha-1}\, y_2 = \frac{\Sigma\, x_1^{2\alpha-1}\, x_2}{n\,(n-1)},$$

and so forth, the sums on the right in the numerators being symmetric functions of the constants $x$, taken over all different terms obtained from that written by permuting subscripts, and the denominator being in each case the number of terms in the numerator. Thus

$$\mu_{2\alpha} = \frac{1}{s_2^{2\alpha}} \left\{ \frac{(\Sigma\, x_1^{2\alpha})^2}{n} + A\, \frac{(\Sigma\, x_1^{2\alpha-1}\, x_2)^2}{n\,(n-1)} + B\, \frac{(\Sigma\, x_1^{2\alpha-2}\, x_2x_3)^2}{n\,(n-1)\,(n-2)} + \cdots \right\}, \tag{13}$$

where the coefficients $A, B, \cdots$ depend on $\alpha$ but not on $n$. With a view to determining the leading term in the expansion of $\mu_{2\alpha}$ in powers of $n^{-1}$, we shall select the term in the curly brackets in (13) of highest degree, meaning by the degree of one of these rational fractions the excess of the degree of the numerator over that of the denominator.

The symmetric functions are well known to be expressible as polynomials in

the power-sums $s_p$. In each term of such a polynomial corresponding to one of our symmetric function of degree $2\alpha$, the sum of the subscripts of the $s_p$'s must be $2\alpha$, since if all the $x$'s are multiplied by a constant such a polynomial must be multiplied by the $2\alpha$th power of the constant. Now $s_p$ is a polynomial of degree $p + 1$ in $n$, if $n$ is even, but vanishes identically if $n$ is odd. Consequently the degree in $n$ of any of the terms of the polynomial in the power-sums must exceed $2\alpha$ by the number of power-sums appearing in this term. Therefore, the term of highest degree in $n$ obtained, when one of the symmetric functions is expressed in terms of the $s_p$'s and thence in terms of $n$, must contain the greatest possible number of the $s_p$'s. If $p$ is the number of distinct $x$'s in a term of one of our symmetric functions, this function may be written in the form

$$
\begin{aligned}
\Sigma x_1^{a_1} x_2^{a_2} \cdots x_p^{a_p} = {} & c_0 s_{a_1} s_{a_2} \cdots s_{a_{p-1}} s_{a_p} - c_1 s_{a_1+a_p} s_{a_2} \cdots s_{a_{p-1}} \\
& - c_2 s_{a_1} s_{a_2+a_p} \cdots s_{a_{p-1}} - \cdots - c_{p-1} s_{a_1} s_{a_2} \cdots s_{a_{p-1}+a_p} \\
& - c' s_{a_1+a_2+a_p} s_{a_3} \cdots s_{a_{p-1}} - \cdots ,
\end{aligned} \tag{14}
$$

where $a_1 + a_2 + \cdots + a_p = 2\alpha$, and the $c$'s do not involve $n$. In the right-hand member of the equation above, the first term involves $p$ of the power-sums, while the remaining terms involve fewer of them. Hence, if all the indices $a_1, a_2, \cdots, a_p$ are even, the first term is a polynomial of degree $2\alpha + p$ in $n$, while the remaining terms are polynomials of lower degree, and are therefore negligible in comparison with the first term when $n$ is sufficiently large. But if any of the indices $a_i$ are odd, the first term vanishes identically, and the degree of (14), regarded as a polynomial in $n$, is then less than $2\alpha + p$. Since the sum of the indices is $2\alpha$, the number of odd ones among them must be even; let this number be denoted by $2q$, and let the number of even indices be $m$. Then $p = m + 2q$. The terms of highest degree in the right-hand member of (14) must be obtained by grouping the odd indices in pairs to form the subscripts of the $s$'s. The degree is therefore $2\alpha + m + q$.

In (13), the degree of the denominator of each term in the curly brackets is the number of distinct $x$'s appearing in a term of the symmetric function in the numerator, namely $p$, or $m + 2q$. Hence the excess of the degree of the numerator over that of the denominator is

$$2\,(2\alpha + m + q) - (m + 2q) = 4\alpha + m\,.$$

This will be a maximum when $m$ is a maximum, and is independent of $q$. The maximum value of $m$ is $\alpha$, and occurs only for the symmetric function

$$\Sigma\, x_1^2 x_2^2 \cdots x_\alpha^2\,. \tag{15}$$

The term involving this function is therefore the only one in the right-hand member of (13) we need consider. Since this symmetric function contains $n(n-1)(n-2)\cdots(n-\alpha+1)/(\alpha!)$ terms, and since in the expansion of

$$(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}$$

the coefficient of $x_1^2x_2^2 \cdots x_\alpha^2 y_1^2 y_2^2 \cdots y_\alpha^2$ is, by the multinomial theorem $(2\alpha)!/2^\alpha$, we have from (13),

$$\mu_{2\alpha} \sim \frac{1}{s_2^{2\alpha}} \frac{(2\alpha)!}{2^\alpha} \frac{\alpha!\,(\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2)^2}{n^\alpha}.$$

To evaluate the symmetric function (15), so far as the term of highest order in, $n$ is concerned, we of course need only the first term of (14), which reduces in this case to

$$\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2 = c_0 s_2^\alpha - \cdots .$$

In the expansion of $s_2^\alpha = (x_1^2 + x_2^2 + \cdots + x_n^2)^\alpha$, the coefficient of (15) is $\alpha!$, which is therefore the reciprocal of $c_0$. Thus we obtain

$$\mu_{2\alpha} = \frac{(2\alpha)!}{\alpha!\,2^\alpha}\left[\frac{1}{n^\alpha} + \cdots\right],$$

the terms dropped being of higher order in $n^{-1}$.

The $2\alpha$th moment of the quotient of $r'$ by its standard error, that is, of $r'\sqrt{n-1}$, is $(n-1)^\alpha$ times that of $r'$, and therefore approaches, as $n$ increases, the value

$$\frac{(2\alpha)!}{\alpha!\,2^\alpha}. \tag{16}$$

The odd moments are all zero because of the symmetry of the distribution of $r'$. But (16) is the moment of order $2\alpha$ of a normal distribution of unit variance and zero mean. It follows therefore from the Second Limit Theorem of Probability[7] that the distribution tends to normality as $n$ increases; that is, for any real number $\lambda$, the limit as $n$ tends to infinity of the probability that $r'\sqrt{n-1} < \lambda$ is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{1}{2}x^2}\,dx .$$

The normality of the limiting distribution of the rank correlation coefficient is rather remarkable, since $r'$, unlike the product-moment correlation coefficient $r$ and other statistics in common use, is neither a mean of independent quantities nor a function of such means, so that the ultimate normality just established is not a corollary of known general theorems. It is unexpected also because the exact distribution of $r'$ for samples smaller than six might lead one to anticipate a bimodal distribution.

An outstanding problem is to determine whether the distribution of $r'$ in samples from a bivariate normal distribution for which $\rho \neq 0$ converges to normality. Without such an approach to normality, the probable error formulae

[7] First proved by Markoff. Cf. Fréchet and Shohat, *A Proof of the Generalized Second Limit Theorem in the Theory of Probability*, Transactions of the American Mathematical Society, Vol. 33, 1932, pp. 533–543.

discovered by Pearson are useless. Another problem is to find convenient and accurate approximations to the distribution of $r'$, for moderate values of $n$, with close limits of error. A table calculated along the lines suggested in §5 would be very useful.

### 8. Combination of Rank and Contingency Methods

Suppose that a thousand school children are examined at the end of a course of instruction, and rated with the grades A, B, C and D. Five hundred of these children are of each sex. The results are:

| | A | B | C | D | Totals |
|---|---|---|---|---|---|
| Boys | 190 | 200 | 80 | 30 | 500 |
| Girls | 220 | 200 | 60 | 20 | 500 |
| Totals | 410 | 400 | 140 | 50 | 1000 |
| Proportion of Girls | .537 | .500 | .429 | .400 | .500 |

Regarding this as a 2 x 4 contingency table with three degrees of freedom, we calculate $\chi^2 = 7.52$, the probability of which value being exceeded by chance is .0570. The indications of a significant difference in distribution of grades between sexes may thus, if one holds to the .05 standard and uses only the $\chi^2$ test, be regarded as not quite significant. There is, however, additional evidence in the fact that the proportion of girls diminishes steadily as we pass down the scale of grades. If we treat excellence in the subject as one variate and the proportion of girls in a group as another, we have a rank correlation of unity, with a sample of four. The probability of a correlation of $\pm 1$ is .083, which also, by itself, would not be considered significant. But we may combine the two pieces of evidence by the method given by Fisher.[8] The process consists of adding the natural logarithms of the two probabilities, doubling, and treating the result as having the $\chi^2$ distribution with four degrees of freedom. This gives a probability in the neighborhood of .03, which would be judged significant.

Similar cases are very common. The value of $\chi^2$ is unchanged if the columns are permuted in any way, whereas $r'$ depends solely on which of the possible permutations actually exists. Thus the two tests are *independent*, a property needed for the combination by the above method.

### 9. Efficiency of Replacement of Measures by Ranks, and the Estimation of $\rho$ from Rank Correlation, for a Normal Population

Consider a population with a normal distribution in two variates $x$ and $y$, each of which we shall without loss of generality assume to be of unit variance and zero mean. The density distribution is then specified by $z\,dx\,dy$, where

[8] R. A. Fisher, *Statistical Methods for Research Workers*, 4th and 5th editions, Art. 21.1.

$$z = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho x y+y^2)}, \tag{17}$$

where $\rho$ is the correlation of $x$ and $y$, or the variate correlation. By $\xi$ and $\eta$, as in §3, we denote the uniformized variates defined by (5), i.e., functions respectively of $x$ and $y$ having distributions of uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$. Then $\xi$ and $\eta$ will each have the variance $1/12$. The rank correlation $\rho'$ in the population is the correlation of $\xi$ and $\eta$; consequently

$$\rho' = 12 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\, \eta\, z\, dx\, dy\,. \tag{18}$$

Thus $\rho'$ is a function of $\rho$, which obviously vanishes when $\rho = 0$.

From (17) the identity

$$\frac{\partial z}{\partial \rho} = \frac{\partial^2 z}{\partial x\, \partial y} \tag{19}$$

is readily calculated. With its help we have from (18) and integrations by parts,

$$\begin{aligned} \frac{d\rho'}{d\rho} &= 12 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta \frac{\partial z}{\partial \rho}\, dx\, dy = 12 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta \frac{\partial^2 z}{\partial x\, \partial y}\, dx\, dy \\ &= 12 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d\xi}{dx}\frac{d\eta}{dy}\, z\, dx\, dy\,. \end{aligned} \tag{20}$$

Now since $x$ and $y$ are normally distributed with unit variance and zero means, the uniformized variates (5) take the form

$$\xi = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt\,, \qquad \eta = \frac{1}{\sqrt{2\pi}} \int_0^y e^{-\frac{t^2}{2}}\, dt\,.$$

Therefore

$$\frac{d\xi}{dx} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad \frac{d\eta}{dy} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Substituting these values and (17) in the last integral in (20) we have,

$$\frac{d\rho'}{d\rho} = \frac{12}{4\pi^2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(2-\rho^2)x^2-2\rho x y+(2-\rho^2)y^2}{2(1-\rho^2)}}\, dx\, dy\,.$$

The double integral, as is well known, equals $\pi$ divided by the square root of the discriminant of the quadratic form in the exponent. This gives

$$\frac{d\rho'}{d\rho} = \frac{6}{\pi\sqrt{4-\rho^2}}\,.$$

Therefore, since $\rho'$ vanishes with $\rho$,

$$\rho' = \frac{6}{\pi} \sin^{-1} \frac{\rho}{2},$$

or

$$\rho = 2 \sin \frac{\pi\rho'}{6}.$$

This is essentially the process used by Pearson.

The last equation suggests that an estimate $r''$ of $\rho$ be based on the rank correlation $r'$ by means of the relation

$$r'' = 2 \sin \frac{\pi r'}{6}.$$

Prefixing a $\delta$ to denote a deviation of sample from population value we have by a Taylor expansion,

$$\delta r'' = \frac{\pi}{3} \cos \frac{\pi\rho'}{6} \delta r' + \cdots,$$

the terms dropped being of higher order in $\delta r'$ than those written, and consequently of higher order in $n^{-1}$. Squaring, taking the expectation, and ignoring the terms of higher order, we have for the case $\rho = \rho' = 0$, by (10),

$$\sigma_{r''}^2 = E(\delta r'')^2 = \frac{\pi^2}{9} \sigma_{r'}^2 = \frac{\pi^2}{9(n-1)},$$

approximately.

The last result enables us to measure the loss of information, at least for large samples, that results from neglecting the exact values of the variates and using only ranks. The product-moment correlation coefficient $r$ has, if $\rho = 0$, the exact variance

$$\frac{1}{n-1},$$

the ratio of which to $\sigma_{r'}^2$ tends as $n$ increases to $9/\pi^2$. Thus the efficiency of the rank correlation method in estimating $\rho$, if $\rho$ is really zero, is $9/\pi^2 = .9119$. This means that the product-moment correlation is approximately as sensitive a test of the existence of a relationship in a normally distributed population with 91 cases as the rank correlation with 100 cases.

The efficiency of $r'$ will of course be different for non-normal populations, and also for normal populations with $\rho \neq 0$. But if the form of the population is known, this knowledge may always be used to supplement the ranks to obtain a more accurate estimate of correlation, or test of relationship. This fact deserves some attention, since a superficial observation of the coincidence of the formula (1)

for the leading term of the variance of an arbitrary uncorrelated population, and the leading term of the formula (10) for the variance of the rank correlation, might suggest that $r'$ is as accurate as $r$. But it may be surmised that the 9 % loss of information found for the bivariate normal distribution is the greatest loss of information in using $r'$ in place of $r$ to test for independence, since for non-normal populations the most efficient estimate of the correlation will not usually be $r$, but a more complicated function of the observations. Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship.

COLUMBIA UNIVERSITY.

# THE ELIMINATION OF PERPETUAL CALENDARS

By John L. Roberts

If we wish to find the day of the week for any date, one way to solve the problem is to use a perpetual calendar. Another way to solve the problem is to calculate the day of the week by mathematical methods. In the past these mathematical methods have been so complicated that it has been much more convenient to use a perpetual calendar. This explains why some people have put themselves to the expense of buying perpetual calendars. The purpose of this article is to provide a mathematical method which is so simple that the entire calculation can be done mentally and which is as convenient as a perpetual calendar. In this article this mathematical method is applied to the Gregorian, Julian, and World calendars. Since a great many records have been made using the Julian and Gregorian calendars, the adoption of the World calendar would not completely eliminate the usefulness of applying the mathematical method to the historical calendars. The mathematical method also shows to what extent the World calendar is a simplification; this is important because proposals to reform the present calendar are attracting world-wide attention.

In the theory of numbers occurs the expression,

$$a \equiv b \bmod p, \tag{1}$$

which is read $a$ is congruent to $b$ modulo $p$, and which means that the difference of $a$ and $b$ is divisible by $p$. Since $p$ in this article is always equal to 7, it is convenient to represent (1) by

$$a \equiv b. \tag{2}$$

Assume $m$ stands for any number which represents any monthday of any month. Assume $w$ stands for any number which represents any day of the week. It is assumed that 7 stands for Sunday, 1 for Monday, 2 for Tuesday, etc. It is assumed that the constant $c$ for any month is the value of $m$ at the first Sunday in that month. Then (2) becomes

$$w \equiv m - c, \tag{3}$$

which enables us to find $w$ if $m$ is known provided the constant $c$ is known for the month in question. Consequently, all we need to complete our theory is to discover a method of finding $c$ for any possible month.

First, there will be discussed rules for finding $c$ for any month of the Gregorian calendar in 1935. An inspection of the calendar shows that $c$ for December is

equal to 1. Since November has 30 days, we can find $c$ for it by adding 2, which is congruent to 30, to the $c$ for December. Since the number of days in September, October, and November is 91, which is congruent to zero, the $c$'s for September and December have the same value. In like manner, since $c$ for September is 1, the $c$ for June is 2, and the $c$ for March is 3. We now have all the theory which is necessary to find $w$ at any date in 1935. For example, suppose we wish to find $w$ for April 17, and know that the $c$ for December is 1. Then, by adding 2 we find that the $c$ for March is 3. We are now in position easily to calculate that the $c$ for April is 7. Applying (3) we find that $w$ at April 17 is 3, which stands for Wednesday.

All that is necessary to complete our theory of the Gregorian calendar is to find rules for finding $c$ for December of any possible year, because, if this is known, we can find $c$ for any month in that year by the method used for 1935. It is convenient to represent the expression, "$c$ for December 1935" by "$C$ for 1935." In like manner $C$ for any calendar year means $c$ for December of that year. Since $C$ for 1935 is 1 and since the number of days in 1936 is 366, which is congruent to 2, subtracting this 2, we find that $C$ for 1936 is 6, because $-1$ is congruent to 6. Knowing $C$ for 1936, we deduce that $C$ for 1940, which is four years later, is 1, because $6 + 2$ is congruent to 1; and that $C$ for 1928 is 2, found by subtracting 4. The $C$'s for 1900, 1928, 1956, and 1984 are equal. Full centuries in order to be leap years must be divisible by 400. Since $C$ for 1900 is 2, we find by adding 1 that $C$ for 2000 is 3. Knowing $C$ for 2000, we deduce by adding 2 that $C$ for 2100 is 5. 1600, 2000, and 2400 have the same value of $C$. If it is assumed that the length of the tropical year is exactly 365.2425 days, we have all the theory which is necessary to find $C$ for any possible year. Although this assumption contains a small error, any further discussion of it would hardly be of any practical interest. The foregoing theory provides complete methods for finding $w$ by means of a series of steps, which are so simple that the entire calculation can be done mentally. For example, suppose we wish to find $w$ for November 29, 1888. Each of the $C$'s for 1800 and 1884 is 7. Therefore, $C$ for 1888 is 2, which is congruent to $7 + 2$. Adding 2, $c$ for November of this year is 4. Applying (3), we find that $w$ at November 29, 1888 is 4, which stands for Thursday. In order to calculate mentally $w$ for any date of the Gregorian calendar, it is only necessary for me to remember the foregoing mathematical method and to remember I was born on November 29, 1888, a Thanksgiving Day.

Deplorable changes were made in the Julian calendar between 45 B.C. and 1 A.D. Also it was not until 325 A.D. that the use of the 7-day week became general throughout the Roman Empire, gradually supplanting the old division of the month into Calends, Nones, and Ides. Therefore, in order to save space, the application of our theory prior to 1 A.D. is left to the reader. Starting with this year it is only necessary to discover a rule for finding the $C$'s of the Julian calendar for the full centuries, because the rules of the Gregorian calendar apply to all other years. October 5, 1582, Old Style was the same day as Oc-

tober 15, 1582, New Style; the Gregorian calendar was born at this date. December 17, 1600, New Style was a Sunday, and was the same day as December 7, 1600, Old Style. Therefore, $C$ for 1600, Old Style is 7. It is now a very simple matter to complete our theory of the Julian calendar. Since $C$ for 1600 is 7, subtracting 1, $C$ for 1500 is 6. 200, 900, and 1600 have the same value of $C$.

In the case of the World calendar the $c$'s for the three months of each of the equal quarters can be found as follows. For the first month $c$ is 1. Therefore, $c$ for the second month is 5, which is congruent to $1 - 3$. Subtracting 2 from this 5, we find that $c$ for the third month is 3.

## NOTES

# ON STANDARD ERROR FOR THE LINE OF MUTUAL REGRESSION

By Y. K. Wong

1. In Pearson's *On Lines and Planes of Closest Fit to System of Points in Space,* he establishes a formula for the mean square residual for the best fitting line in $q$-space:

$$(1) \qquad (\text{mean sq. residual})^2 = \sigma_{x_1}^2 + \cdots + \sigma_{x_q}^2 - \Delta R_{\max}^2$$

where $2R_{\max}$ is the length of the maximum axis of the correlation ellipse in $q$-space, and $\Delta$ is the correlation determinant.[1]

In the present paper, we consider a 2-dimensional case, and shall call the mean sq. residual as the standard error, denoted by $S_N$.

In 2-dimensional space, a correlation ellipse is

$$(2) \qquad ax^2 + 2hxy + by^2 + c = 0,$$

where

$$(2\text{a}) \quad a = \sigma_y^2, \quad b = \sigma_x^2, \quad h = -r_{xy}\sigma_x\sigma_y = -p_{xy} = -p_{yx}, \quad c = -\sigma_x^2\sigma_y^2.$$

Pearson gives in the 2-dimensional space the following formula for $S_N$:

$$(3) \qquad S_N = \sigma_x\sigma_y/\text{semi-major axis of equation (2)}.$$

Expression (3) can be readily deduced from (1). This paper aims to present some formulae for $S_N$, more convenient for practical computation, and also call attention to a misprint in Pearson's paper.

2. From analytic geometry, we see that the angle $\varphi$, between the major axis of the ellipse (2) and the $x$-axis is given by

$$(4) \qquad \tan 2\varphi = 2h/(a - b).$$

By rotation of the axes, equation (1) can be written in the form

$$(5) \qquad a'x^2 + b'y^2 + c = 0,$$

where

$$(5\text{a}) \quad \begin{aligned} a' &= a\cdot\cos^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\sin^2\varphi > 0 \\ b' &= a\cdot\sin^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\cos^2\varphi > 0. \end{aligned}$$

[1] Philosophical Magazine, 6th Series, II (November, 1901), p. 559.

LEMMA 1. The value of $a'$ given by (5a) is less than $b'$.

To prove this lemma, we find from (4) and (5)

$$a' - b' = a + b, \qquad a' - b' = 2h/\sin 2\varphi = -2p_{xy}/\sin 2\varphi,$$

and hence

$$(6) \qquad 2a' = a + b - 2p_{xy}/\sin 2\varphi, \qquad 2b' = a + b + 2p_{xy}/\sin 2\varphi.$$

Since both $a$ and $b$ are positive, the lemma will be proved if we can show that $p_{xy}/\sin 2\varphi$ is a positive quantity. By (2a), $p_{xy} = r_{xy}\sigma_x\sigma_y$, in which $\sigma_x$, $\sigma_y$ are positive; hence the sign of $p$ depends upon the sign of $r$. If $r_{xy} < 0$, then $\varphi > \frac{\pi}{2}$, and $2\varphi$ is of such a nature that $\frac{3\pi}{2} < 2\varphi < 2\pi$. It follows $\sin 2\varphi < 0$, and hence $p_{xy}/\sin 2\varphi$ is positive. On the other hand, if $r_{xy} > 0$, then $\varphi$ is such that $0 < 2\varphi < \pi$, and hence $\sin 2\varphi > 0$. It follows that $p_{xy}/\sin 2\varphi$ is positive independent of the sign of $r_{xy}$.

LEMMA 2. The square of the mean square residual is equal to $a'$, and hence

$$S_N^2 = \sigma_y^2 \cos^2 \varphi - 2p_{xy} \sin \varphi \cos \varphi + \sigma_x^2 \sin^2 \varphi = \tfrac{1}{2}(\sigma_x^2 - \sigma_y^2) - p_{xy}/\sin 2\varphi.$$

For from (5), we obtain (semi-major axis)$^2 = -c/a' = +\dfrac{\sigma_x^2\sigma_y^2}{a'}$. Substituting this into (3), we obtain $S_N = a'$. The balance of the lemma follows from (5a), (6), and (2a).

LEMMA 3. For every $r_{xy}$, we have

$$(7) \qquad \sin 2\varphi = p_{xy}/\sqrt{K}, \qquad K = (\sigma_x^2 - \sigma_y^2)^2 + 4p_{xy}^2.$$

For, from (4), we find $\sin 2\varphi = -p_{xy}/\pm\sqrt{K} = r_{xy}\left(\dfrac{-\sigma_x\sigma_y}{\pm\sqrt{K}}\right)$. By the argument given in the demonstration of Lemma 1, we see that $r_{xy}$ and $\sin 2\varphi$ should be of the same sign. Hence the negative sign is chosen before the radical.

From Lemma 2 and (7), we have the formula given by Pearson:

$$(8) \qquad 2S_N^2 = (\sigma_x^2 + \sigma_y^2)^2 - \sqrt{K}.$$

3. We are going to establish several more formulae for $S_N$. From (4), we have $2h \cdot \tan(\varphi) = -(a - b) \pm \sqrt{K}$. The sign before the radical is determined in such a way that $\tan(\varphi)$ has the same sign as $r_{xy}$. By the reasoning given in Lemma 1, the negative sign is chosen. Thus

$$-2p_{xy} \cdot \tan \varphi = -(\sigma_y^2 - \sigma_x^2) - \sqrt{K} = \sigma_x^2 + \sigma_y^2 - \sqrt{K} - 2\,\sigma_y^2$$

or

$$2(\sigma_y^2 - p_{xy} \tan \varphi) = \sigma_x^2 - \sigma_y^2 - \sqrt{K}.$$

This proves that

$$(9) \qquad S_N^2 = \sigma_y^2 - p_{xy} \tan \varphi.$$

Similarly, we have

$$(10) \qquad S_N^2 = \sigma_x^2 - p_{xy} \cot \varphi .$$

For computation, (9) and (10) are more convenient than (8). When the line of mutual regression is determined, it is known that $\tan \varphi$ (denoted by $B$) is equal to the slope of that line, and hence $\cot \varphi$ ($= 1/B$) is equal to the reciprocal of the slope. Then we can write (9) and (10) as follows:

$$(11) \qquad S_N^2 = \sigma_y^2 - p_{yx} \cdot B$$

$$(12) \qquad S_N^2 = \sigma_x^2 - p_{xy}/B .$$

The second formula given in Lemma 2 is simpler than (8), but not as simple as (11) and (12).

For computation, it is convenient to find $\varphi$ from the equation

$$\tan 2\varphi = \frac{+2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} = H ,$$

i.e.,

$$2\,\varphi = \arctan H .$$

Since $\sin 2\varphi$ and $r_{xy}$ are of the same sign, we can determine the value of $\varphi$ from the preceding equation by inspection, though $\arctan H$ is a multiple-valued function. After the determination of $\varphi$, we can obtain

$$B = \tan \varphi .$$

Then we can compute $S_N$ either from (9), (11), or (10), (12).

There is a very interesting fact furnished by (11) and (12). These two formulae are, in fact, generalizations of the following two well known ones:

$$(a) \qquad S_y^2 = \sigma_y^2(1 - r)$$

$$(b) \qquad S_x^2 = \sigma_x^2(1 - r) ,$$

where $S_y$ is the standard error of the line of regression when $y$ is used as dependent variable and $x$ as independent variable, and similarly for $S_x$. It is clear that the line of mutual regression may be looked upon as a generalization of the other two lines of regression when we use $y$ or $x$ as dependent variable. So the slope $B$ of the line of mutual regression is a generalization of $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$. where the subscript $yx$ means $y$ on $x$ and $xy$, $x$ on $y$. If we use $x$ as independent variable, then we must obtain $b_{yx}$ instead of $B$. Hence substituting the formula of $b_{yx}$ instead of $B$ into (11), we obtain, after a simple reduction, the same result as given by (a). On the other hand, if we use $y$ as independent variable, we must obtain $b_{xy}$ instead of $1/B$. It will result (b) when $b_{xy}$ is put in the place of $1/B$ in (12). The generalization perhaps can be seen more clearly if we write (a) and (b) into slightly different forms:

(a′) $$S_y^2 = \sigma_y^2 - p_{yx} \cdot b_{yx}$$

(b′) $$S_x^2 = \sigma_x^2 - p_{xy} \cdot b_{xy}.$$

4. The misprint in Pearson's paper is on the second formula of the following:

$$(MSR)^2 = \frac{\sigma_x^2 \sigma_y^2}{\cot^2 \varphi} = \frac{1}{2}\left(\sigma_x^2 - \sigma_y^2 - \sqrt{(\sigma_x^2 - \sigma_y^2)^2 - 4r^2 \sigma_x^2 \sigma_y^2}\right)$$

where $\tan 2\varphi = 2r_{xy}\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)$. $\cot^2 \varphi$ should read "square of semi-major axis of ellipse (2)." Professor Henry Schultz first noticed this misprint and suggested to the writer to investigate it.

In a recent letter to Schultz, Pearson pointed out that one of the simplest formula for $S_N^2$ or $(MSR)^2$ is given by

($\alpha$) $$S_N^2 = \sigma_x^2 \sin^2 \varphi + \sigma_y^2 \cos^2 \varphi,$$

where $\varphi$ is defined by (4). However, Professor Schultz expressed doubt about its validity. From lemma 2, it is clear that ($\alpha$) is also not true.

INSTITUTE OF SOCIAL SCIENCES,
ACADEMIA SINICA, PEIPING

# THE DISTRIBUTION LAWS OF THE DIFFERENCE AND QUOTIENT OF VARIABLES INDEPENDENTLY DISTRIBUTED IN PEARSON TYPE III LAWS[1]

By Solomon Kullback

Although the results herein described are not entirely new, it is felt that the method of solution is of interest as presenting further illustrations of the application of characteristic functions to the distribution problem of statistics (*1*).

**1. Distribution law of the difference.** Let $u = x - y$, where the distribution laws of $x$ and $y$ are independent and given respectively by

$$f_1(x) = \frac{e^{-x}\, x^{p-1}}{\Gamma(p)}\, ; \quad f_2(y) = \frac{e^{-y}\, y^{q-1}}{\Gamma(q)} \qquad 0 \leqq x \leqq \infty\, ; 0 \leqq y \leqq \infty. \tag{1}$$

The characteristic function of the distribution law of $u$ is given by (*1*),

$$\varphi(t) = \int_0^\infty \frac{e^{itx-x}\, x^{p-1}\, dx}{\Gamma(p)} \int_0^\infty \frac{e^{-ity-y}\, y^{q-1}\, dy}{\Gamma(q)} \tag{2}$$

$$= \frac{1}{(1 - it)^p\, (1 + it)^q}. \tag{3}$$

The distribution law of $u$ is given by (*1*),

$$D(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itu}\, dt}{(1 - it)^p\, (1 + it)^q}. \tag{4}$$

Let $1 - it = - \dfrac{z}{u}$,

$$D(u) = \frac{e^{-u}\, u^{p-1}}{2^q\, 2\pi i} \int_{-u-i\infty}^{-u+i\infty} \frac{e^{-z}\, dz}{(-\, z)^p \left(1 + \dfrac{z}{2u}\right)^q}. \tag{5}$$

Now it may be shown that (*1*)

$$\frac{1}{2\pi i} \int_{-u-i\infty}^{-u+i\infty} \frac{e^{-z}\, dz}{(-\, z)^p \left(1 + \dfrac{z}{2u}\right)^q} = \frac{e^{u}\, (2u)^{\frac{q-p}{2}}}{\Gamma(p)}\, W_{\frac{p-q}{2}, \frac{1-p-q}{2}}\, (2u) \tag{6}$$

---

[1] Presented to the American Mathematical Society, June 20, 1934.

where $W_{k,m}(z)$ is the confluent hypergeometric function (*2*). Since $W_{k,m}(z) = W_{k,-m}(z)$ we have finally

$$(7) \qquad D(u) = \frac{u^{\frac{p+q}{2}-1}}{2^{\frac{p+q}{2}}\,\Gamma(p)}\, W_{\frac{p-q}{2},\,\frac{p+q-1}{2}}(2u).$$

For $p = q$, since $W_{0,m}(2x) = \dfrac{x^{\frac{1}{2}}\,2^{\frac{1}{2}}}{\sqrt{\pi}} K_m(x)$ where $K_m(x)$ is the Bessel Function of the second kind and imaginary argument (*1*), we obtain

$$(8) \qquad D(u) = \frac{u^{\frac{2p-1}{2}}}{2^{\frac{2p-1}{2}}\,\Gamma(p)\sqrt{\pi}}\, K_{\frac{2p-1}{2}}(u).$$

This result has been otherwise obtained by Pearson, Stouffer, and David (*3*).

**2. Distribution law of the quotient.** Let $u = \log x - \log y$ where $x$ and $y$ are defined as above.

The characteristic function of the distribution law of $u$ is given by (*1*)

$$(9) \qquad \varphi(t) = \int_0^\infty \frac{e^{-x}\,x^{p-1+it}\,dx}{\Gamma(p)} \int_0^\infty \frac{e^{-y}\,y^{q-1-it}\,dy}{\Gamma(q)}$$

$$(10) \qquad = \frac{\Gamma(p+it)\,\Gamma(q-it)}{\Gamma(p)\,\Gamma(q)}.$$

The distribution law of $u$ is given by (*1*)

$$(11) \qquad D(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itu}\,\Gamma(p+it)\,\Gamma(q-it)}{\Gamma(p)\,\Gamma(q)}\,dt.$$

Let $q - it = -z$, so that

$$(12) \qquad D(u) = \frac{e^{-qu}}{\Gamma(p)\,\Gamma(q)\,2\pi i}\int_{-q-i\infty}^{-q+i\infty} e^{-zu}\,\Gamma(p+q+z)\,\Gamma(-z)\,dz.$$

Now it may be shown that (*2*)

$$\frac{1}{2\pi i}\int_{-q-i\infty}^{-q+i\infty} e^{-zu}\,\Gamma(p+q+z)\,\Gamma(-z)\,dz = \Gamma(p+q)\,(1+e^{-u})^{-(p+q)},$$

so that

$$(13) \qquad D(u) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\,\frac{e^{pu}}{(1+e^{u})^{p+q}}.$$

Since $e^u = \frac{x}{y} = w$, we obtain as the distribution law of the quotient

$$p(w) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)} \frac{w^{p-1}}{(1+w)^{p+q}}. \tag{14}$$

If in (13) we set

$$p = \frac{n_1}{2}; \qquad q = \frac{n_2}{2}; \qquad e^u = \frac{n_1}{n_2} e^{2z},$$

we obtain

$$D(z) = \frac{2\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} e^{n_1 z}}{(n_2 + n_1 e^{2z})^{\frac{n_1+n_2}{2}}} \tag{15}$$

which result has been otherwise obtained by R. A. Fisher (*4*).

George Washington University.

REFERENCES

(1) Kullback, S.: An application of characteristic functions to the distribution problem of statistics. Annals of Mathematical Statistics, Vol. 5 (1934) pp. 263–307.

(2) Whittaker and Watson: Modern Analysis, 2nd Ed., pp. 333, 283.

(3) Pearson, Stouffer and David: Further applications in statistics of the Bessel Function. Biometrika, Vol. 24 (1932), pp. 293.

(4) Fisher, R. A.: On a distribution yielding the error functions of several well known statistics. Proceedings International Mathematical Congress, Toronto (1924), Vol. 2, pp. 805–813.

## REPORT OF THE MEETING OF THE INSTITUTE OF MATHEMATICAL STATISTICS AT ST. LOUIS

The Institute of Mathematical Statistics held a joint meeting with the Econometric Society and the American Mathematical Society at St. Louis, Missouri, on January 2, 1936. The program consisted of an invited address, "The Mathematical Theory of Index Numbers," by Professor Thomas Rawles, and the following additional papers:

(1) On Certain Distributions Derived from the Multinomial Distribution, by Dr. Solomon Kullback

(2) Convexity Properties of Generalized Mean Value Functions, by Dr. Nilan Norris

(3) The Frequency Distribution for the Mean of $n$ Independent Chance Variables When Each Is Subject to the Law $y_0 x^{p-1}(1-x)^{q-1}$, by Prof. W. D. Baten

(4) On the Admissibility of Time Series, by Prof. Francis Regan

The Institute voted to hold a meeting at Cambridge, Massachusetts, early in September of this year. This meeting will be in connection with the celebration of the Harvard Tercentenary. Professor R. A. Fisher will deliver an invited address before the Institute and the American Mathematical Society. A more detailed announcement of the meeting will be made later.